EXHIBIT 11

5 Econometric Models of Probabilistic Choice

Daniel McFadden

An object can have no value unless it has utility. No one will give anything for an article unless it yield him satisfaction. Doubtless people are sometimes foolish, and buy things, as children do, to please a moment's fancy; but at least they think at the moment that there is a wish to be gratified.
—F. M. Taussig, *Principles of Economics*, 1912

5.1 Economic Man

The classical economists made the assumption of *homus economicus* virtually tautological: if an object were chosen, then it must have maximized the utility of the decision maker. By contrast, contemporary economic analysis of consumer behavior has focused on the objective market environment of economic decisions and has excluded whim and perception from any formal role in the utility maximization process.[1]

From the standpoint of the observer unmeasured psychological factors introduce a random element in economic decisions. The result is a probabilistic theory of choice which has many features in common with psychophysical models of judgment (Coombs 1964, Luce and Suppes 1965, Bock and Jones 1968, Krantz, Luce, Suppes, and Tversky 1971, Krantz 1974).

Probabilistic choice models lend themselves readily to econometric implementation, particularly for choices among discrete alternatives. This chapter develops and compares a number of these models in forms suitable for econometric applications.

This research was supported in part by the National Science Foundation, through grant SOC75-22657 to the University of California, Berkeley. Portions of this chapter were written while the author was an Irving Fisher Visiting Professor of Economics at the Cowles Foundation for Research in Economics, Yale University. An early version was presented at the Third World Congress of the Econometric Society, Toronto, Canada, 1975. The author has benefited greatly from discussions with Amos Tversky at the formative stage of this chapter and has borrowed freely from his ideas. Charles Manski and Steven Cosslett have also provided useful comments. The author retains sole responsibility for errors.

1. Also excluded in the conventional consumer analysis is consideration of procedural rationality, the question of how an organism with perceptual and computational limits makes a decision; see Simon (1978). This chapter will not take up the question of probabilistic choice theory in the presence of bounded rationality. However, we note that the distributions of demand attributed in this chapter to taste variation or errors in judgment could often be reinterpreted as a consequence of bounded rationality, and vice versa.

5.2 Discrete Choice

Many empirically important economic decisions involve choice among discrete alternatives. Examples are decisions on labor force participation, occupation, educational level, marital status, family size, residential and work location, travel mode, and brands of commodity purchases. The problem of economic discrete choice parallels the decision context in which psychophysical models have been applied successfully. On the other hand, analysis of discrete choice behavior using conventional marginalist consumer theory is quite awkward. For these reasons we concentrate on a probabilistic consumer theory for discrete choice.

An example helps to clarify the conceptual and empirical issues involved in the study of discrete choice. Consider the choice by commuters of auto or bus mode to work. For the example assume the number of commuters is fixed, so that we can concentrate on the proportion of commuters choosing bus. We expect this proportion to be a function of the relative travel times and costs of the two modes. An empirical approach to forecasting, say, the effect of transit fares on bus patronage, would be to fit a demand function to aggregate time-series data, disregarding theoretical foundations. However, a priori information on the form and structure of the demand function implied by an analysis of decision behavior may permit sharper forecasts. In particular, if the relationship between individual decisions and aggregate demand is understood, then extensive data on individual choices can be used to refine estimates of the aggregate demand function.

An approach to such problems often used by the new home economists is to assume that individual demand is the result of utility maximization by a representative consumer whose decision variable is the proportion of trips taken by bus. Since this decision variable is continuous, conventional marginal analysis applies. Market demand is pictured as the aggregate of a population of identical representative consumers, so that market demand is just individual demand writ large.[2]

While the single representative consumer model may be a useful analytic device under appropriate assumptions (see section 5.6), it provides a poor description of individual behavior. What we observe is a population split into mostly full-time auto users or full-time bus users. The effect of rising

2. See, for example, Becker and Stigler (1977), where the conceptual foundation for common tastes is advocated with scant attention to the practicalities of econometric demand analysis with limited data on consumer characteristics.

transit fares is felt primarily at the extensive margin where some individuals are switching from bus to auto. The conceptual implausibility of a model of identical consumers with fractional consumption rates is even more obvious for decisions such as level of education or family size.

These comments on the difficulty of using the concept of identical representative consumers as a basis for modeling discrete choice behavior provide the kernel of a solution for the problem. Suppose we introduce a population of consumers in which tastes vary explicitly. For example, we might consider a population of consumers with quadratic utility functions whose coefficients are distributed in the population according to some specific parametric probability distribution. Then, we can express proportions such as the share of bus commuters in our example as the probability that an individual drawn randomly from the population will have tastes such that the utility of traveling by bus exceeds that for auto. The parameters of the aggregate demand function will then be the parameters of the underlying probability distribution of taste coefficients.

The idea of taste variation in a population influencing aggregate demand behavior is an old one. Many of the classical consumer demand studies, such as Prais and Houthakker (1971), discuss this as a nuisance to be eliminated by assumption. Seminal studies by Tobin (1958), Warner (1962), and Quandt (1968) use the idea explicitly in the analysis of specific problems of discrete and limited choice. More recently the analysis of econometic models with random parameters has been motivated by the presence of unobserved variations among economic agents. The topics discussed in this chapter are a natural extension of the idea of taste variation to general questions in discrete choice analysis.

The demand behavior of populations of consumers can be analyzed at two distinct levels. At a theoretical level we can examine the general implications for the structure of aggregate demands that can be drawn from the hypothesis of a population of preference-maximizing consumers. The most basic question is whether individual preference maximization has any implications for aggregate demand structure. A related question is whether the model of individual utility maximization is identifiable from the observed distributions of demands, or whether other simpler or less restrictive models could generate the same observations. These topics are discussed in sections 5.3 through 5.8. These sections also consider sufficient conditions for an aggregate demand system to be consistent with the hypothesis of a population of preference-maximizing consumers.

At a practical level we can take as given the hypothesis of a population of preference maximizers and seek parametric demand structures suitable for econometric analysis. Sections 5.9 through 5.16 survey a number of alternative model structures and summarize their features. Sections 5.17 through 5.19 treat in more detail the estimation of a proposed model structure, the tree extreme value, or nested multinomial logit model. Appendixes give computational formulae for several of the models.

5.3 Probabilistic Consumer Theory

In sections 5.4 through 5.8 we shall first define a probabilistic choice system, describing the observable distributions of demands by a population of consumers. Second, we shall state the hypothesis of random preference maximization, which postulates that the distribution of demands in a population is the result of individual preference maximization, with preferences influenced by unobserved variables. Third, we consider the features of the observable distributions of demands that are necessary or sufficient for their consistency with the hypothesis of random preference maximization.

The development of population demand behavior parallels exactly the conventional treatment of the individual consumer, with distributions of observed demands and preferences replacing a single demand system and preference order. The usual necessary conditions for consistency of an individual demand system with preference maximization have population analogues, as does a stochastic version of the theory of revealed preference. Sufficient (integrability) conditions for an observed demand distribution to be consistent with a distribution of preferences are much less complete than the analogous treatment of individual demand.

5.4 Probabilistic Choice Systems

A probabilistic choice system (PCS) is defined formally by a vector $(\mathbf{I}, \mathbf{Z}, \xi, \mathscr{B}, \mathbf{S}, P)$, where $\mathbf{I}$ is a set indexing alternatives, $\mathbf{Z}$ is the universe of vectors of measured attributes of alternatives, $\xi : \mathbf{I} \to \mathbf{Z}$ is a mapping specifying the observed attributes of alternatives, $\mathscr{B}$ is a family of finite, nonempty choice (or budget) sets from $\mathbf{I}$, $\mathbf{S}$ is the universe of vectors of measured characteristics of individuals, and $P : \mathbf{I} \times \mathscr{B} \times \mathbf{S} \to [0,\ 1]$ is a choice probability.

The index set $\mathbf{I}$ is imposed by the analyst and is assumed to be external to the actual choice process. Any natural or intrinsic indexing of alternatives which may affect choice is included in the vector of measured attributes $\mathbf{z} \in \mathbf{Z}$. The universe of measured attributes $\mathbf{Z}$ will be treated here as an abstract set; in later applications it will usually be assumed to be a rectangle, or else a countable dense set, in finite-dimensional Euclidean space. The choice probability $P(i \mid \mathbf{B}, \mathbf{s})$ specifies the probability of choosing $i \in \mathbf{I}$, given that a selection must be made from the choice set $\mathbf{B} \in \mathscr{B}$ and that the decision-maker has characteristics $\mathbf{s} \in \mathbf{S}$. We use the notation $P(\mathbf{C} \mid \mathbf{B}, \mathbf{s}) = \Sigma_{i \in \mathbf{C}} P(i \mid \mathbf{B}, \mathbf{s})$. Choice probabilities are assumed to satisfy the following two conditions:

PCS 5.1: Choice probabilities are nonnegative and sum to one, with $P(\mathbf{B} \mid \mathbf{B}, \mathbf{s}) = 1$.

PCS 5.2: Choice probabilities depend only on the measured attributes of alternatives and individual characteristics; if $\mathbf{B} = \{i_1, \ldots, i_n\}$ and $\mathbf{B}' = \{i'_1, \ldots, i'_n\}$ have $\mathbf{z}_k = \xi(i_k) = \xi(i'_k)$ for $k = 1, \ldots, n$, then $P(i_k \mid \mathbf{B}, \mathbf{s}) = P(i'_k \mid \mathbf{B}', \mathbf{s})$.

It should be noted that a PCS is analogous to a conventional econometric specification of a demand system, with the functional specification of the demand structure and the distribution of errors combined to specify the distribution of demand.

5.5 The Random Utility Maximization Hypothesis

The hypothesis of random utility maximization (RUM) is defined formally by a vector $(\mathbf{I}, \mathbf{Z}, \xi, \mathbf{S}, \mu)$, where $(\mathbf{I}, \mathbf{Z}, \xi, \mathbf{S})$ are defined as for a PCS, and μ is a probability measure depending on $\mathbf{s} \in \mathbf{S}$, on the space of utility functions on $\mathbf{I}$.[3] The probability μ gives the distribution of tastes in the population of individuals with characteristics $\mathbf{s} \in \mathbf{S}$.[4]

3. The space of utility functions is $\mathbf{R}^{\mathbf{I}}$, where $\mathbf{R}$ is the real line. Give $\mathbf{R}^{\mathbf{I}}$ the product topology, and define the measurable sets in $\mathbf{R}^{\mathbf{I}}$ to be the Borel sets in this topology.

4. Each utility function is a specified ordinal representation of a preference relation on $\mathbf{I}$. One could alternately start from a probability measure η on the set of preference relations on $\mathbf{I}$. From this random preference maximization model, choice probabilities could be deduced directly. When the preference relations are representable by utility functions, the measure η on preferences and the representation mapping induce a measure μ on the space of utility functions. Technically a preference relation on $\mathbf{I}$ is defined as a subset ρ of $\mathbf{I} \times \mathbf{I}$ containing all the pairs (i, j) with i at least as desirable as j, and having the properties that $(i, j) \in \rho$ and $(j, k) \in \rho \Rightarrow (i, k) \in \rho$. Let $\mathbf{T}$ be the set of all

Let $\mathscr{B}$ denote a family of nonempty, finite choice sets, as earlier. Let $\mu^{\mathbf{B}}$ denote the restriction of μ to $\mathbf{B} \in \mathscr{B}$.[5] The following assumptions are imposed on μ:

RUM 5.1: The restriction of μ to the space of utility values on a finite set of alternatives $\mathbf{B} \in \mathscr{B}$ depends on the measured attributes of these alternatives; if $\mathbf{B} = \{i_1, \ldots, i_n\}$ and $\mathbf{B}' = \{i'_1, \ldots, i'_n\}$ have $\mathbf{z}_k = \xi(i_k) = \xi(i'_k)$ for $k = 1, \ldots, n$, then $\mu^{\mathbf{B}} = \mu^{\mathbf{B}'}$.

RUM 5.2: The probability of "ties" is zero;

$\mu(\{\mathbf{U} \in \mathbf{R}^{\mathbf{I}} \mid u(i_1) = u(i_2)\}, \mathbf{s}) = 0.$

The next assumption states that choice is determined by utility maximization.

RUM 5.3: Each RUM ($\mathbf{I}$, $\mathbf{Z}$, ξ, S, μ) and family of choice sets $\mathbf{B} \in \mathscr{B}$ generates a PCS ($\mathbf{I}$, $\mathbf{Z}$, ξ, $\mathscr{B}$, S, P) via the following mapping: for $\mathbf{B} = \{i_1, \ldots, i_n\} \in \mathscr{B}$, $\mathbf{s} \in \mathbf{S}$, and $k = 1, \ldots, n$,

$$P(i_k \mid \mathbf{B}, \mathbf{s}) = \mu(\{\mathbf{U} \in \mathbf{R}^{\mathbf{I}} \mid U(i_k) \geqq U(i_j) \text{for} j = 1, \ldots, n\}, \mathbf{s}). \tag{5.1}$$

The assumption RUM 5.2 guarantees that there is almost always a unique utility-maximizing alternative, so that (5.1) is well defined, with $P(\mathbf{B} \mid \mathbf{B}, \mathbf{s}) = 1$.

When the restriction of μ to $\mathbf{B} = \{i_1, \ldots, i_n\} \in \mathscr{B}$ can be represented by a probability density $f^{\mathbf{B}}$, so that $\mu^{\mathbf{B}}(\mathbf{A}, \mathbf{s}) = \int_{\mathbf{A}} f^{\mathbf{B}}(u_1, \ldots, u_n; \mathbf{s})\, du_1 \ldots du_n$ for each measurable subset $\mathbf{A}$ of $\mathbf{R}^n$, then the choice probabilities can be rewritten

preference relations on $\mathbf{I}$, and $\mathscr{I}$ a Boolean σ-algebra of subsets of $\mathbf{T}$. Then $\eta: \mathscr{I} \times \mathbf{S} \to [0, 1]$ is a probability measure provided $\eta(\cdot, \mathbf{s})$ is nonnegative and countably additive on $\mathscr{I}$, with $\eta(\mathbf{T}) = 1$. Suppose a subset $\mathbf{T}_0 \subseteq \mathbf{T}$ is measurable and has $\eta(\mathbf{T}_0) = 1$, and that there exists a measurable mapping $\psi: \mathbf{T}_0 \to \mathbf{R}^{\mathbf{I}}$, giving an ordinal representation of each $\rho \in \mathbf{T}_0$. The probability measure μ will obviously depend on the choice of the representation mapping ψ. In many applications $\mathbf{I}$ can be assumed countable. Then every preference relation on $\mathbf{I}$ has a representation $U: \mathbf{I} \times \mathbf{T} \to [0, 1]$ defined by $U(i, \rho) = \bigcup_{j \in \mathbf{A}(i, \rho)} 2^{-j}$, where $\mathbf{A}(i, \rho) = \{j \in \mathbf{I} \mid (i, j) \in \rho\}$.
More general representation theorems are discussed in Debreu (1962). Note that the range of ordinal utility can be restricted without loss of generality to the unit interval, so that all positive moments can be assumed to exist.

5. For $\mathbf{B} = \{i_1, \ldots, i_n\} \in \mathscr{B}$, $\mu^{\mathbf{B}}$ is a probability measure on the finite-dimensional space $\mathbf{R}^n$ of vectors $(u(i_1), \ldots, u(i_n))$ of utility levels for the alternatives in $\mathbf{B}$.

$$P(i_1 \mid \mathbf{B},\mathbf{s}) = \int_{u_1=-\infty}^{+\infty} \int_{u_2=-\infty}^{u_1} \cdots \int_{u_n=-\infty}^{u_1} f^{\mathbf{B}}(u_1, \ldots, u_n; \mathbf{s})\, du_1 \ldots du_n$$

$$= \int_{u=-\infty}^{+\infty} F_1^{\mathbf{B}}(u, \ldots, u; s)\, du, \tag{5.2}$$

where $F^{\mathbf{B}}$ is the cumulative distribution function of $f^{\mathbf{B}}$, and $F_1^{\mathbf{B}}$ denotes the derivative of $F^{\mathbf{B}}$ with respect to its first argument. Alternately, letting $G^{\mathbf{B},1}(w_2, \ldots, w_n; \mathrm{s})$ denote the cumulative distribution function of $(w_2, \ldots, w_n) = (u(i_2) - u(i_1), \ldots, u(i_n) - u(i_1))$, the choice probability satisfies[6]

$$P(i_1 \mid \mathbf{B},\mathbf{s}) = G^{\mathbf{B},1}(0, \ldots, 0; \mathbf{s}). \tag{5.3}$$

The problem of finding econometrically feasible PCS consistent with RUM is attacked by using (5.2) to generate choice probabilities constructively from parametric families of probabilities μ, or by demonstrating constructively or indirectly that candidate PCS are consistent with some probability μ.

5.6 Stochastic Revealed Preference

Does the hypothesis of a population of utility-maximizing consumers imply any restrictions on the distributions of observed demands? An affirmative answer was given by Marschak (1960) and Block and Marschak (1960), who established the necessity of conditions such as regularity and the triangle inequality.[7] A necessary and sufficient condition for consistency with random preference maximization, analogous to the strong axiom of revealed preference for the individual consumer, has been established by McFadden and Richter (1970). Let $(\mathbf{B}, \mathbf{C})$ be a pair of sets with $\mathbf{B} \in \mathscr{B}$ and $\mathbf{C} \subseteq \mathbf{B}$. If an individual offered an alternative from $\mathbf{B}$ makes

6. The relation of $F^{\mathbf{B}}$ and $G^{\mathbf{B},1}$ is

$$G^{\mathbf{B},1}(w_2, \ldots, w_n) = \int_{u=-\infty}^{+\infty} F_1^{\mathbf{B}}(u, u + w_2, \ldots, u + w_n)\, du.$$

7. A PCS satisfies regularity if $\mathbf{C} \subseteq \mathbf{B} \subseteq \mathbf{B}' \Rightarrow P(\mathbf{C} \mid \mathbf{B}, \mathrm{s}) \geqq P(\mathbf{C} \mid \mathbf{B}', \mathrm{s})$, and the triangle inequality if $P(i \mid \{i, j\}, \mathrm{s}) \leqq P(i \mid \{i, k\}, \mathrm{s}) + P(k \mid \{k, j\}, \mathrm{s})$.

a selection in **C**, call (**B**, **C**) a successful trial. Then the strong axiom of revealed stochastic preference states that for any finite sequence of trials $(\mathbf{B}^1, \mathbf{C}^1), \ldots, (\mathbf{B}^M, \mathbf{C}^M)$, with repetitions permitted,

$$\sum_{m=1}^{M} P(\mathbf{C}^m \mid \mathbf{B}^m, \mathbf{s}) \leqq N((\mathbf{B}^1, \mathbf{C}^1), \ldots, (\mathbf{B}^M, \mathbf{C}^M)), \tag{5.4}$$

where $N((\mathbf{B}^1, \mathbf{C}^1), \ldots, (\mathbf{B}^M, \mathbf{C}^M))$ is the maximum number of successful trials in the sequence consistent with a single preference order. This axiom implies a variety of necessary conditions that can be used to screen PCS for consistency;[8] however, it does not provide a practical sufficient condition for consistency.

Suppose a PCS is consistent with RUM. Are there alternative theories of individual behavior which can generate the same PCS, but which for reasons of generality are to be preferred to the classical model of individual utility maximization? One more general alternative is immediate. We might view the individual himself as drawing a utility function from a random distribution each time a decision is made. Then the individual is a classical utility maximizer given his state of mind, but his state of mind varies randomly from one choice situation to the next.[9] Intraindividual and interindividual variations in tastes are indistinguishable in their effect on the observed distribution of demand.

The hypothesis of intrapersonal random utility is appealing on methodological grounds, since it fits the same data as the conventional theory,

8. Consider, for example, the PCS known as the maximum model (McFadden 1974) with $\mathbf{I} = \{1, 2, 3, 4\}$. Let $\mathbf{I}_1 = \{1, 2\}$ and $\mathbf{I}_2 = \{3, 4\}$. The binary choice probabilities satisfy $P(i \mid ij) = v_i/(v_i + v_j)$, with $v_1 = 3$, $v_2 = 2$, $v_3 = 4$, $v_4 = 3$ for the example. For choice sets of more than two alternatives, only the available alternative in $\mathbf{I}_1$ with the highest scale value is retained, and similary for $\mathbf{I}_2$, with choice between the retained alternatives satisfying the binary choice probabilities; for example, $P(1 \mid 123) = v_1/(v_1 + v_3)$ and $P(2 \mid 123) = 0$. For the trials (12, 1), (34, 3), (234, 2), and (124, 4), equation (5.4) yields $P(1 \mid 12) + P(3 \mid 34) + P(2 \mid 234) + P(4 \mid 124) = 421/210 > 2 =$ the maximum number of successes consistent with RUM. Hence the maximum model can fail to satisfy the axiom of revealed stochastic preference.

9. We confine our attention to the case where the drawings of utility functions for successive decisions are independent. More generally one could introduce learning, experience, and habit by making the probability distribution over utility functions dependent on history. Data collected on series of decisions by cross sections of individuals would permit the identification of intraindividual and interindividual components of variation in tastes. Some of the econometric analysis for this extension is given in Heckman, chapters 3 and 4. A general treatment of the topic awaits future research.

with weaker postulates.[10] The intrapersonal random utility model is in fact of considerable historical and contemporary importance in psychological theories of individual choice. It was first suggested by Thurstone (1927), and it forms the basis for many current models of individual choice behavior put forward in psychology by Luce (1959), Tversky (1972), and others, and tested with reasonable success. In addition to providing evidence on the plausibility of the intrapersonal random utility model as a theory of individual choice behavior, the psychological literature provides analytic results and functional forms that can be adapted for economic applications.

5.7 Aggregation of Preferences

One useful method for examining the consistency of PCS with RUM is to test compatibility with sufficient conditions for consistency. The author is unaware of any general analogue for RUM of the simple sufficient (integrability) conditions of individual utility theory. A restricted, but useful, result of this sort is obtained when individual preferences have sufficient structure to aggregate to a social (indirect) utility function yielding aggregate demands. In this case the home economist's traditional representative consumer with fractional consumption rates can be assigned the social utility function, justifying this approach as an analytic shortcut consistent with some underlying population of utility maximizers who make discrete choices.

Suppose the consumption of an individual is defined by a vector $\mathbf{x}$ of the quantities consumed of divisible commodities and choice of a discrete alternative i which has a vector of measured intrinsic attributes $\mathbf{w}$. The individual has a utility function $\tilde{U}:\tilde{\mathbf{X}} \times \mathbf{W} \times \mathbf{I} \rightarrow [0, 1]$, where $\tilde{\mathbf{X}} \times \mathbf{W}$ is the space of pairs of vectors $(\mathbf{x}, \mathbf{w})$. The utility function is assumed to satisfy the direct utility (DU) assumption

DU: $\tilde{\mathbf{X}}$ is the nonnegative orthant of a finite-dimensional real vector space, and $\mathbf{W}$ is a closed set in a finite-dimensional real vector space. The utility function $\tilde{U}(\cdot, \cdot, i)$ is continuous on $\tilde{\mathbf{X}} \times \mathbf{W}$ for each $i \in \mathbf{I}$. $\tilde{U}(\cdot, \mathbf{w}, i)$ is twice continuously differentiable on $\tilde{\mathbf{X}}$, with $\partial\tilde{U}/\partial\mathbf{x} \geqq 0$ and

10. A particular attraction is that the hypothesis permits retention of much of the apparatus of classical welfare economics. If a criterion for interpersonal comparisons exists in the theory, then it can be applied to intrapersonal comparisons as well.

$|\partial\tilde{U}/\partial\mathbf{x}| > 0$, and is strictly differentially quasi-concave, for each $\mathbf{w} \in \mathbf{W}$ and $i \in \mathbf{I}$.[11]

The individual has income y and faces a vector of prices $\mathbf{r} \gg \mathbf{0}$ for divisible commodities and a cost q associated with the discrete alternative. For a specified discrete alternative i with measured attributes $\mathbf{w}$, the individual chooses $\mathbf{x}$ to maximize utility subject to the budget constraint $\mathbf{r}\cdot\mathbf{x} + q = y$. The result is a conditional indirect utility function $V(y-q, \mathbf{r}, \mathbf{w}, i; \tilde{U})$ defined for $y - q > 0$, $\mathbf{r} \gg \mathbf{0}$, $\mathbf{w} \in \mathbf{W}$, $i \in \mathbf{I}$, and $\tilde{U}$ satisfying assumption DU by

$$V(y-q,\mathbf{r},\mathbf{w},i;\tilde{U}) = \max_{\mathbf{x}} \ \{\tilde{U}(\mathbf{x},\mathbf{w},i) \mid \mathbf{r}\cdot\mathbf{x} \leqq y - q\}. \tag{5.5}$$

This function has the indirect utility (IU) properties:[12]

IU 5.1: For $\mathbf{r} \gg \mathbf{0}$, $y - q > 0$, $\mathbf{w} \in \mathbf{W}$, $i \in \mathbf{I}$, and $\tilde{U}$ satisfying the utility conditions of DU $V(y-q, \mathbf{r}, \mathbf{w}, i; \tilde{U})$ is continuous in $(y - q, \mathbf{r}, \mathbf{w})$, twice continuously differentiable and homogeneous of degree zero in $(y - q, \mathbf{r})$, strictly differentiably quasi-convex in $\mathbf{r}$, and has $\partial V/\partial(y-q) > 0$.[13]

IU 5.2: (Roy's identity): The maximum of $\tilde{U}(\mathbf{x}, \mathbf{w}, i)$ subject to $\mathbf{r}\cdot\mathbf{x} \leqq y - q$ is achieved at a unique vector $\mathbf{x} = \mathbf{X}(y - q, \mathbf{r}, \mathbf{w}, i; \tilde{U})$ which satisfies

$$\mathbf{X}(y-q,\mathbf{r},\mathbf{w},i;\tilde{U}) = -\frac{\partial V/\partial\mathbf{r}}{\partial V/\partial y}. \tag{5.6}$$

When $\mathbf{r} \gg \mathbf{0}$ and $y - q > 0$ are confined to a compact set, there exists a monotone transformation of $\tilde{U}$, given by by $\hat{U} = (e^{\alpha U} - 1)/(e^{\alpha} - 1)$, which for α sufficiently large implies the corresponding transformation of V is convex in $\mathbf{r}$.[14] Thus for most applications V can be assumed without further loss of generality to be convex in $\mathbf{r}$.

Suppose the consumer faces a finite set of discrete alternatives $\mathbf{B} \in \mathscr{B}$. With alternative $i \in \mathbf{B}$ is associated a vector of measured attributes, $\mathbf{z}_i =$

11. $\tilde{U}(\mathbf{x})$ is strictly differentiably quasi-concave if $\mathbf{t}\cdot\partial\tilde{U}/\partial\mathbf{x} = 0$ and $\mathbf{t}\cdot\mathbf{t} = 1$ imply $\mathbf{t}'[\partial^2\tilde{U}/\partial\mathbf{x}\partial\mathbf{x}']\mathbf{t} < 0$; see McFadden (1978b, pp. 30, 368).

12. See, particularly, Diewert (1977), and also McKenzie (1957) and McFadden (1978b, p. 34).

13. Strict differential quasi-convexity requires if $\mathbf{t}\cdot\partial V/\partial\mathbf{r} = 0$ and $\mathbf{t}\cdot\mathbf{t} = 1$, then $\mathbf{t}'(\partial^2 V/\partial\mathbf{r}\partial\mathbf{r}')\mathbf{t} > 0$.

14. The hessian of $\hat{V} = (e^{\alpha V} - 1)/(e^{\alpha} - 1)$ is $\hat{V}_{\mathbf{rr}} = (V_{\mathbf{rr}} + \alpha V_{\mathbf{r}} V_{\mathbf{r}}')\alpha e^{\alpha V}/(e^{\alpha} - 1)$, where $V_{\mathbf{rr}} \equiv \partial^2 V/\partial\mathbf{r}\partial\mathbf{r}'$. A sequence $(y - q, \mathbf{r})_k$ in the compact set, $\alpha_k \to +\infty$, $\mathbf{t}_k \cdot V_{\mathbf{r}} = 0$, $\mathbf{t}_k \cdot \mathbf{t}_k = 1$, and $\mathbf{t}_k'\hat{V}_{\mathbf{rr}}\mathbf{t}_k \leqq 0$ has a subsequence converging to $\mathbf{t}^*$ and $(y - q, \mathbf{r})_*$ at which $\mathbf{t}^*\cdot V_{\mathbf{r}} = 0$ and $\mathbf{t}^{*\prime}V_{\mathbf{rr}}\mathbf{t}^* \leqq 0$, contradicting the strict differential quasi-convexity of V in $\mathbf{r}$. Hence there exists a finite positive α for which the result holds.

$(q_i, \mathbf{r}, \mathbf{w}_i) = \xi(i)$ in our earlier terminology. Income y is a component of the vector s of consumer characteristics. The unconditional indirect utility function of the consumer is then

$$V^*(y - q_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}; \tilde{U}) = \max_{i \in \mathbf{B}} V(y - q_i, \mathbf{r}, \mathbf{w}_i, i, \tilde{U}), \tag{5.7}$$

where $y - \mathbf{q_B}$ denotes a vector with a component $y - q_j$, and $\mathbf{w_B}$ a vector with a component w_j, for each $j \in \mathbf{B}$. For almost all $y - \mathbf{q_B}$, consumer demand for the discrete alternatives is given by Roy's identity,[15]

$$\delta_j = D(j \mid \mathbf{B}, \mathrm{s}; \tilde{U}) \equiv -\frac{\dfrac{\partial V^*}{\partial q_j}}{\dfrac{\partial V^*}{\partial y}} \equiv \begin{cases} 1 & \text{if } j \in \mathbf{B} \text{ and } v_j \geqq v_k \text{ for } k \in \mathbf{B}, \\ 0 & \text{otherwise}, \end{cases} \tag{5.8}$$

where $v_k = V(y - q_k, \mathbf{r}, \mathbf{w}_k, k; \tilde{U})$. The population choice probabilities then satisfy[16]

$$\begin{aligned} P(j \mid \mathbf{B}, \mathrm{s}) &= E_{U|\mathrm{s}} D(j \mid \mathbf{B}, \mathrm{s}; \tilde{U}) \\ &= \int D(j \mid \mathbf{B}, \mathrm{s}; \tilde{U}) \mu(d\tilde{\mathbf{U}}, \mathrm{s}) \\ &= \mu(\{\tilde{U} \in \mathbf{R}^{\mathbf{I}} \mid V(y - q_j, \mathbf{r}, \mathbf{w}_j, j; \tilde{U}) \\ &\qquad \geqq V(y - q_k, \mathbf{r}, \mathbf{w}_k, k; \tilde{U}) \text{ for } k \in \mathbf{B}\}, \mathrm{s}). \end{aligned} \tag{5.9}$$

We shall now seek sufficient conditions on preferences such that a social utility function can be defined, with fractional consumption rates for the discrete alternatives, which yields the PCS (5.9). This problem is closely related to the classical analysis of community preferences by Gorman (1953); see also Chipman (1974), Muellbauer (1976), Shapiro (1975), and Lau (1977).

15. Except for $\mathbf{q_B}$ in a closed set of Lebesgue measure zero and (by RUM 5.2) μ-probability zero, the maximum in (5.7) is achieved at a unique alternative. On the open set where $k \in \mathbf{B}$ is the unique maximum, $V^*(y - \mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}; \tilde{U}) = V(y - q_k, \mathbf{r}, \mathbf{w}_k; \tilde{U})$, and V^* shares the regularity and differentiability properties of V given in IU 5.1, for almost all price vectors.

16. Given $(q_i, \mathbf{r}, \mathbf{w}_i) = \xi(i)$ and the maximizing vector $\mathbf{X}(y - q_i, \mathbf{r}, \mathbf{w}_i, i; \tilde{U})$, one obtains for each i the conditional indirect utility $U(i; \tilde{U}) \equiv \tilde{U}(\mathbf{X}(\xi(i), i; \tilde{U}), i) \equiv V(\xi(i), i; \tilde{U})$. Thus there is a mapping from the set of $\tilde{U}$ to the set of $U: \mathbf{I} \to [0, 1]$. With a slight abuse of notation, we write $\mu(d\tilde{U}, \mathrm{s})$ rather than $\mu(dU, \mathrm{s})$, with the understanding that the mapping above is applied.

First, we define a utility function with fractional consumption rates. Define $\Delta = \{\boldsymbol{\delta} \in \mathbf{R}^{\mathbf{I}} \mid \delta_i \geqq 0 \text{ and } \Sigma_i \delta_i = 1\}$ and $\Delta_{\mathbf{B}} = \{\boldsymbol{\delta} \in \Delta \mid \delta_i = 0 \text{ for } i \notin \mathbf{B}\}$ for $\mathbf{B} \in \mathscr{B}$. Consider $\bar{U}: \tilde{\mathbf{X}} \times \Delta \times \mathbf{S} \to [0, 1]$. For $\mathbf{r} \gg \mathbf{0}$, $\mathbf{B} \in \mathscr{B}$, $y - \mathbf{q}_{\mathbf{B}} \gg \mathbf{0}$, and $\mathbf{w}_{\mathbf{B}} \in \mathbf{W}$, with $(y - q_i, \mathbf{r}, \mathbf{w}_i) = \xi(i)$, define

$$\begin{aligned}&\bar{V}(y - q_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) \\ &= \max_{\mathbf{x}, \boldsymbol{\delta}} \{\bar{U}(\mathbf{x}, \boldsymbol{\delta}, \mathbf{s}) \mid \mathbf{x} \in \mathbf{X}, \boldsymbol{\delta} \in \Delta_{\mathbf{B}}, \mathbf{r} \cdot \mathbf{x} + \mathbf{q}_{\mathbf{B}} \cdot \boldsymbol{\delta}_{\mathbf{B}} \leqq y\}. \qquad (5.10)\end{aligned}$$

We term $\bar{U}$ a social utility function, and $\bar{V}$ a social indirect utility function, if the choice probabilities satisfy Roy's identity,

$$P(i \mid \mathbf{B}, \mathbf{s}) = -\frac{\dfrac{\partial \bar{V}}{\partial q_i}}{\dfrac{\partial \bar{V}}{\partial y}}. \qquad (5.11)$$

Suppose individual conditional indirect utility functions have the form

$$V(y - q, \mathbf{r}, \mathbf{w}, i; \tilde{U}) = \frac{y - q - \alpha(\mathbf{r}, \mathbf{w}, i; \tilde{U})}{\beta(\mathbf{r})}, \qquad (5.12)$$

where $y > q + \alpha(\mathbf{r}, \mathbf{w}, i; \tilde{U})$ and α and β are homogeneous of degree one, concave, and nondecreasing in $\mathbf{r}$.[17] The linearity of V in $(y - q)$ implies that V^* in (5.7) is additively separable into a term independent of $\tilde{U}$ and a term independent of y. Consider the function $\bar{V}$ defined by[18]

$$\begin{aligned}\bar{V}(y - \mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) &= E_{U \mid \mathbf{s}} \max_{i \in \mathbf{B}} V(y - q_i, \mathbf{r}, \mathbf{w}_i, i; \tilde{U}) \\ &\equiv \frac{1}{\beta(\mathbf{r})}\left\{y + E_{U \mid \mathbf{s}} \max_{i \in \mathbf{B}} [-q_i - \alpha(\mathbf{r}, \mathbf{w}_i, i; \tilde{U})]\right\}. \qquad (5.13)\end{aligned}$$

17. That (5.12) is an indirect utility function follows immediately from the concavity of the associated expenditure function, $y = q + \alpha(\mathbf{r}, \mathbf{w}, i; \tilde{U}) + u\beta(\mathbf{r})$, for $u \geqq 0$. The quasi-convexity of V in $(y - q, \mathbf{r})$ can also be demonstrated by a direct calculus argument. The aggregation properties of (5.12) were first noted by Gorman (1953), who provided the following characterization of the direct preference map: $\tilde{U}(\mathbf{x}, \mathbf{w}, i) = \operatorname{Max}_{\hat{\mathbf{x}}} \{U^0(\hat{\mathbf{x}}) \mid U^1(\mathbf{x} - \hat{\mathbf{x}}, \mathbf{w}, i) = 1\}$, where U^0 and U^1 are concave in $\hat{\mathbf{x}}$, U^1 is homogeneous of degree zero in $\mathbf{x} - \hat{\mathbf{x}}$, and U^0 does not vary over the population. This dual structure can also be derived from composition rules for concave conjugate functions; see McFadden (1978b, p. 49–60).

18. It is assumed here that the expectation exists. Note however that, while the ordinal utility function $\tilde{U}$ can be assumed to have a range contained in the unit interval, and thus have an expectation, the transformation of utility necessary to achieve additive separability in (5.12) may yield a function whose expectation (5.13) does not exist. In section 5.8, a modified definition of $\bar{V}$ is employed which precludes this possibility.

The terms $[-q_i - \alpha(\mathbf{r}, \mathbf{w}, i; \tilde{U})]$ are convex in $(\mathbf{q_B}, \mathbf{r})$. Since the maximum of convex functions is convex, and a nonnegative linear combination of convex functions is convex,

$$G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) \equiv E_{U\,|\,\mathbf{s}} \max_{i \in \mathbf{B}} \, [-q_i - \alpha(\mathbf{r}, \mathbf{w}_i, i; \tilde{U})] \tag{5.14}$$

is convex in $(\mathbf{q_B}, \mathbf{r})$. Then $\bar{V} = (y + G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})/\beta(\mathbf{r})$ is invertible to a concave expenditure function $y = u\beta(\mathbf{r}) - G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})$ for $u \geqq 0$ and is therefore an indirect utility function.

Applying (5.8) to this preference structure yields

$$D(j\,|\,\mathbf{B}, \mathbf{s}; \tilde{U}) = -\frac{\partial}{\partial q_j} \max_{i \in \mathbf{B}} \, [-q_i - \alpha(\mathbf{r}, \mathbf{w}_i, i; \tilde{U})], \tag{5.15}$$

and hence from (5.9)

$$\begin{aligned} P(j\,|\,\mathbf{B}, \mathbf{s}) &= E_{U\,|\,\mathbf{s}} D(j\,|\,\mathbf{B}, \mathbf{s}; \tilde{U}) \\ &\equiv -\frac{\partial G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})}{\partial q_j} \\ &\equiv -\frac{\partial \bar{V}/\partial q_j}{\partial \bar{V}/\partial y} \quad . \end{aligned} \tag{5.16}$$

Therefore $\bar{V}$ is a social indirect utility function yielding the PCS.[19]

When this conclusion holds, the demand distribution can be analyzed as if it were generated by a population with common tastes, with each (representative) consumer having fractional consumption rates for the discrete alternatives and the social indirect utility function $\bar{V}$.

It should be noted that the utility structure (5.12) yields choice probabilities that are independent of current income. However, tastes (the distribution of $\tilde{U}$) may depend on individual characteristics that are correlates of current income such as historical wage rates, income levels, or occupation. Then these variables may enter the PCS.

5.8 The Williams-Daly-Zachary Theorem

The conclusion derived from the preference structure (5.12), that the resulting choice probabilities are given by the gradient of a surplus function

19. The associated direct utility function $\bar{U}$ satisfies

$$\bar{U}(\mathbf{x}, \boldsymbol{\delta}, \mathbf{s}) = \operatorname*{Inf}_{y, \mathbf{q}, \mathbf{r}, \mathbf{B}} \{\bar{V}(y - \mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) \,|\, \mathbf{r} \cdot \mathbf{x} + \mathbf{q_B} \cdot \boldsymbol{\delta}_\mathbf{B} \leqq y, \, \mathbf{B} \in \mathscr{B}\}.$$

G satisfying (5.14) and (5.16), can be strengthened by giving necessary and sufficient conditions on G for (5.14) and (5.16) to hold. These conditions will then provide practical criteria for the derivation of PCS consistent with RUM having the structure (5.12). The essential elements of the following arguments are due to Williams (1977) and to Daly and Zachary (1976).

Consider a preference structure satisfying RUM and representable in the additively separable form (5.12). For $\mathbf{B} \in \mathscr{B}$, let $F(\boldsymbol{\varepsilon}_{\mathbf{B}}; \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ denote the cumulative distribution function, induced by the probability measure on the set of $\tilde{U}$, of the random vector $\boldsymbol{\varepsilon}_{\mathbf{B}}$ with components $\varepsilon_i = -\alpha(\mathbf{r}, \mathbf{w}_i, i; \tilde{U})$ for $i \in \mathbf{B}$. If F can be characterized by a density $f(\boldsymbol{\varepsilon}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$, then this random preference structure will be said to be of additive income random utility maximizing, AIRUM, form.[20]

A function $G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ will be termed a social surplus, SS, function if it has the following properties:

SS 5.1: For $\mathbf{B} = \{1, \ldots, m\} \in \mathscr{B}$, G is a real-valued function of $\mathbf{q}_{\mathbf{B}} \in \mathbf{R}^m$, $\mathbf{r} \in \tilde{\mathbf{X}}$ with $\mathbf{r} \geqslant \mathbf{0}$, $\mathbf{w}_{\mathbf{B}} \in \mathbf{W}^m$, and $\mathbf{s} \in \mathbf{S}$.

SS 5.2: G is a positively linear homogeneous, convex function of $(\mathbf{q}_{\mathbf{B}}, \mathbf{r})$.

SS 5.3: G has the additivity property that $G(\mathbf{q}_{\mathbf{B}} + \theta, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) - \theta$, where θ is any real scalar and $\mathbf{q}_{\mathbf{B}} + \theta$ denotes a vector with components $q_i + \theta$.

SS 5.4: All mixed partial derivatives of G with respect to $\mathbf{q}_{\mathbf{B}}$ exist, are nonpositive and independent of the order of differentiation, and satisfy $G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) - G(\mathbf{0}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = \int_0^1 (d/dt) G(\psi(t), \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})\, dt$, where ψ is any path between $\psi(0) = \mathbf{0}_{\mathbf{B}}$ and $\psi(1) = \mathbf{q}_{\mathbf{B}}$.[21]

SS 5.5: $\lim_{q_i \to -\infty} G_i(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = -1$ for $i \in \mathbf{B}$.

SS 5.6: Suppose $\mathbf{B} = \{i_1, \ldots, i_m\} \in \mathscr{B}$, and $\mathbf{B}' = \{i'_1, \ldots, i'_m, \ldots, i'_{m+n}\} \in \mathscr{B}$ satisfy $(q_{i_k}, \mathbf{w}_{i_k}) = (q_{i'_k}, \mathbf{w}_{i'_k})$ for $k = 1, \ldots, m$. Then $G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = G((\mathbf{q}_{\mathbf{B}'}, +\infty, \ldots, +\infty), \mathbf{r}, \mathbf{w}_{\mathbf{B}'}, \mathbf{B}', \mathbf{s})$.

20. The condition for F to be characterized by a density is that it be absolutely continuous with respect to Lebesgue measure on $\mathbf{R}^m$. Note that the linear homogeneity of $\alpha(\mathbf{r}, \mathbf{w}_i, i; \tilde{U})$ in $\mathbf{r}$ implies that $F(\lambda\boldsymbol{\varepsilon}_{\mathbf{B}}, \lambda\mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = F(\boldsymbol{\varepsilon}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ for $\lambda > 0$. When there is no ambiguity, the abbreviated notation $F(\boldsymbol{\varepsilon}_{\mathbf{B}})$ and $f(\boldsymbol{\varepsilon}_{\mathbf{B}})$ will be used. Note that in the utility structure (5.12), β can in general be a function of $\mathbf{r}$ and $\mathbf{s}$.

21. The partial derivative of a function G with respect to its ith argument is denoted G_i. Then $G_{1,2,\ldots,m}$ denotes the mixed partial derivative of G with respect to $(q_1, \ldots, q_m)$. ($\mathbf{0}_{\mathbf{B}}$ is an m-vector of zeroes.)

Consider a PCS with choice probabilities given by functions $P(i \mid \mathbf{B}, \mathbf{s}) = \pi_i(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})$. This will be termed a translation-invariant probabilistic choice system, TPCS, if it satisfies PCS and the following conditions:

TPCS 5.1: The functions π_i are defined for $i \in \mathbf{B} = \{1, \ldots, m\} \in \mathscr{B}$, $\mathbf{q_B} \in \mathbf{R}^m$, $\mathbf{r} \in \tilde{\mathbf{X}}$ with $\mathbf{r} \gg \mathbf{0}$, $\mathbf{w} \in \mathbf{W}^m$, and $\mathbf{s} \in \mathbf{S}$.

TPCS 5.2: π_i is homogeneous of degree zero in $(\mathbf{q_B}, \mathbf{r})$.

TPCS 5.3: For a real scalar θ, $\pi_i(\mathbf{q_B} + \theta, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) = \pi_i(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})$.

TPCS 5.4: $\lim_{q_i \to -\infty} \pi_i(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) = 1$.

TPCS 5.5: All mixed partials of π_i with respect to components of $\mathbf{q_B}$ other than q_i exist, are nonnegative and independent of order of differentiation, and satisfy

$$\pi_1(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})$$
$$= \int_{-\infty}^{q_2} \cdots \int_{-\infty}^{q_m} \pi_{1,2,\ldots,m}(q_1, \hat{q}_2, \ldots, \hat{q}_m, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})\, d\hat{q}_2, \ldots, d\hat{q}_m,$$

with analogous conditions for $\pi_2, \ldots, \pi_m$.[22]

TPCS 5.6: $\pi_{i,j}(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) = \pi_{j,i}(\mathbf{q_B}, \mathbf{r}. \mathbf{w_B}, \mathbf{B}, \mathbf{s})$.

TPCS 5.7: Suppose $\mathbf{B} = \{i_1, \ldots, i_m\} \in \mathscr{B}$ and $\mathbf{B}' = \{i'_1, \ldots, i'_m, \ldots, i'_{m+n}\} \in \mathscr{B}$ satisfy $(q_{i_k}, \mathbf{w}_{i_k}) = (q_{i'_k}, \mathbf{w}_{i'_k})$ for $k = 1, \ldots, m$. Then, for $k = 1, \ldots, m$,

$$\pi_k(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) = \pi_k((\mathbf{q}'_{\mathbf{B}'}, +\infty, \ldots, +\infty), \mathbf{r}, \mathbf{w}_{\mathbf{B}'}, \mathbf{B}', \mathbf{s}).$$

The following theorem links additive-income random utility-maximizing forms, social surplus functions, and translation-invariant probabilistic choice systems.

THEOREM 5.1: Consider $\mathbf{B} = \{1, \ldots, m\} \in \mathscr{B}$.

1. Suppose AIRUM holds, with individual indirect utility having the form $u(i) = (y - q_i + \varepsilon_i)/\beta(\mathbf{r})$, with $\boldsymbol{\varepsilon}_{\mathbf{B}}$ distributed in the population with

22. The partial derivative of π_i with respect to its jth argument is denoted $\pi_{i,j}$. Then $\pi_{1,2,\ldots,m}$, denotes the mixed partial derivative of π_1 with respect to $(q_2, \ldots, q_m)$.

cumulative distribution function $F(\boldsymbol{\varepsilon}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ and density $f(\boldsymbol{\varepsilon}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$. Define

$$G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = \int_{t=-\infty}^{+\infty} [F(\mathbf{0}_{\mathbf{B}} + t, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) - F(\mathbf{q}_{\mathbf{B}} + t, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})]\, dt. \tag{5.17}$$

Then G exists and is a social surplus function satisfying SS.
Further

$$\bar{V}(y - \mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = (y + G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}))/\beta(\mathbf{r}) \tag{5.18}$$

is a social indirect utility function; that is, the PCS associated with this AIRUM form satisfies

$$P(i \mid \mathbf{B}, \mathbf{s}) \equiv \pi_i(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = -G_i(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) \tag{5.19}$$

and satisfies TPCS.

2. Suppose $G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ is a social surplus function satisfying SS. Then (5.19) defines a PCS satisfying TPCS. Further there exists an AIRUM form such that G satisfies (5.17) and (5.18).

3. Suppose $P(i \mid \mathbf{B}, \mathbf{s}) = \pi_i(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ is a PCS, satisfying TPCS. Then there exist an AIRUM form and a social surplus function that satisfy SS and (5.17) through (5.19).

LEMMA 5.1: If AIRUM holds, and the distribution $F(\boldsymbol{\varepsilon}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ has first moments, then G defined by (5.17) also equals

$$G(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s}) = \int_{-\infty}^{+\infty} \left\{ \max_{i \in \mathbf{B}} (\varepsilon_i - q_i) - \max \varepsilon_i \right\} f(\boldsymbol{\varepsilon}_{\mathbf{B}})\, d\boldsymbol{\varepsilon}_{\mathbf{B}}. \tag{5.20}$$

Because the definition (5.17) of the social surplus function normalizes its value to zero for $\mathbf{q}_{\mathbf{B}} = \mathbf{0}_{\mathbf{B}}$, for any nonprice attributes $\mathbf{w}_{\mathbf{B}}$, the social indirect utility function (5.18) cannot be used to make welfare comparisons when nonprice attributes change. The following result gives a modified definition of the social surplus function which permits such comparisons.

LEMMA 5.2: Suppose AIRUM holds. If the distribution $F(\mathbf{q}_{\mathbf{B}}, \mathbf{r}, \mathbf{w}_{\mathbf{B}}, \mathbf{B}, \mathbf{s})$ has first moments, then

$$G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) = \int_{-\infty}^{+\infty} \max_{i \in \mathbf{B}} (\varepsilon_i - q_i) f(\boldsymbol{\varepsilon}_\mathbf{B}) d\boldsymbol{\varepsilon}_\mathbf{B}$$

is a social surplus function, satisfying SS, which when substituted in (5.18) yields a social indirect utility function, permitting welfare comparisons for nonprice attribute changes.

Alternately suppose nonprice attributes are compensable in the sense that given $\mathbf{w_B}$, $\delta > 0$, there exists $\theta > 0$ such that $F(\boldsymbol{\varepsilon}_\mathbf{B} + \theta, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) \geqq F(\boldsymbol{\varepsilon}_\mathbf{B}, \mathbf{r}, \mathbf{w}'_\mathbf{B}, \mathbf{B}, \mathrm{s}) \geqq F(\boldsymbol{\varepsilon}_\mathbf{B} - \theta, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s})$ for all $\boldsymbol{\varepsilon}_\mathbf{B}$ and $\mathbf{w}'_\mathbf{B}$ with $| \mathbf{w}'_\mathbf{B} - \mathbf{w_B} | < \delta$. Then $G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) = \int_{-\infty}^{+\infty} [F(\mathbf{0_B} + t, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}) - F(\mathbf{q_B} + t, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s})] dt$, where $\bar{\mathbf{w}}_\mathbf{B}$ is a fixed vector of nonprice attributes, is a social surplus function satisfying SS. When this function is substituted in (5.18), the social indirect utility function permits welfare comparisons for a fixed set $\mathbf{B}$ and subsets of $\mathbf{B}$ (choice sets formed by letting $q_i \to +\infty$ for some $i \in \mathbf{B}$), and for nonprice attribute changes.

The proofs of the theorem and lemmas are lengthy and are deferred to section 5.23. Several comments on this theorem are in order. First, the conditions TPCS are usually easy to check for an empirical PCS. If they hold, the PCS is consistent with RUM. Thus TPCS is a useful set of sufficient conditions for consistency. Note, however, that, while TPCS is necessary and sufficient for an AIRUM form, there are many PCS consistent with RUM that fail to satisfy TPCS and AIRUM. Second, a useful way to generate PCS consistent with RUM is to start from a social surplus function satisfying SS. A variety of functional forms are known that satisfy SS; several are given in the remainder of this chapter.

A third comment is on welfare analysis of alternative policies involving discrete choice. When preferences have an AIRUM form, the social indirect utility function (5.18), incorporating a social surplus function defined by (5.17) or by lemma 5.2, permits ready comparison of the social desirability of alternative policies. When the vector of prices $\mathbf{r}$ of nondiscrete commodities is unchanged under alternative policies, welfare comparisons can also be made using the social surplus function G. Then G yields an analytic expression for Hicksian consumer's surplus and (since income effects are absent) Marshallian consumer's surplus: for any path $\boldsymbol{\psi}:[0, 1] \to \mathbf{R}^m$ with $\psi(0) = \mathbf{0_B}$, $\boldsymbol{\psi}(1) = \mathbf{q_B}$,

$$G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) \equiv -\int_{t=0}^{1} \sum_{i=1}^{m} \pi_i(\psi(t), \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})\psi_i'(t)\,dt, \tag{5.21}$$

the usual sum of areas under demand curves. Since consistent use of consumer surplus welfare comparisons is grounded on utility structures of the additively separable form (5.12) in conventional problems, we conclude that the presence of discrete choice places no new restrictions on the validity of consumer surplus methods. For further discussion of consumer welfare judgments involving discrete choice, see Rosen and Small (1979).

A fourth comment regards the definition of the variables $\mathbf{q_B}$. In the argument surrounding (5.12), q_i was interpreted as the price of alternative i. The utility function (5.12) was interpreted as the indirect utility function resulting from maximization of a translated homothetic utility function subject to a budget constraint. However, (5.12) can alternately be given the interpretation of a utility function that is additively separable and linear in some physical attribute of the discrete alternative. Consider a utility function with the general structure of (5.12),

$$u = -\frac{q_i + \alpha(\mathbf{r}, \mathbf{w}_i, i, \mathbf{s}; \tilde{U})}{\beta(\mathbf{r}, \mathbf{s})}. \tag{5.22}$$

Take q_i to be the level of some physical attribute of alternative i, and $\mathbf{w}_i$ to be a vector of the remaining attributes of the alternative, including its price. Let α and β depend on the vector s of individual characteristics, including income. Then theorem 5.1 can be applied to establish the existence of a social surplus function G and probabilistic choice system π_i which satisfy (5.17), (5.19), and all the conditions SS and TPCS except the homogeneity properties SS 5.2 and TPCS 5.2. With suitable added assumptions on α and β, (5.21) will be an indirect utility function satisfying IU, and hence will be dual to a direct utility function satisfying DU.[23] This interpretation permits a very general dependence of the PCS on income and prices. However, the logic of the interpretation requires that it be sensible to consider alternatives in which the attribute q_i varies, with all other attributes remaining unchanged. The additively separable utility structure

23. An example that satisfies IU 5.1 for all real values of q_i is $\beta(\mathbf{r}, \mathbf{s})$ a (positive) constant and α a quasi-concave, zero-degree homogeneous function of $\mathbf{r}$ and the income component y of s which is twice continuously differentiable in $(\mathbf{r}, y)$, has $\partial\alpha/\partial y < 0$ and $\partial\alpha/\partial\mathbf{r} \geqq 0$, and is strictly differentiably quasi-concave.

also requires that marginal rates of substitution between attributes other than q_i not depend on the level of q_i.

The reinterpretation of (5.22) can also be made for noneconomic choice contexts, with (**r**, **s**) interpreted as individual characteristics, (q_i, $\mathbf{w}_i$) as physical attributes of the alternative, and u as the direct utility associated with the alternative.

Finally, note that q_i may itself be a function of underlying raw attributes of the alternatives. This function may be parametric; however, it cannot depend on tastes $\tilde{U}$.

Suppose ($\mathbf{L}$, $\mathscr{L}$, J) is a probability space, and G^j is a social surplus function and π_i^j the associated PCS, for each $j \in \mathbf{L}$. Then it is obvious that the probability mixture $G^* = \int_{\mathbf{L}} G^j J(dj)$ is again a social surplus function, with a probabilistic choice system given by the corresponding probability mixture $\pi_i^* = \int_{\mathbf{L}} \pi_i^j J(dj)$. This observation can be used to derive a variety of PCS obtained as mixtures of simpler PCS.[24]

Suppose $\pi_i(\mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})$ is an arbitrary probabilistic choice system depending on individual and economic characteristics (**s**, **r**) and alternative attributes $\mathbf{w_B}$. Then the choice system

$$\tilde{\pi}_i(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) = \frac{e^{-q_i}\pi_i(\mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})}{\sum_{j \in \mathbf{B}} e^{-q_j}\pi_j(\mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})}, \tag{5.23}$$

where the $\mathbf{q_B}$ are artificial variables, satisfies all the conditions TPCS except the homogeneity condition TPCS 5.2 and the condition TPCS 5.7 that the choice probabilities of a set depend only on the measured attributes of the alternatives in that set. The proof of theorem 5.1 then implies the existence of a social surplus function

$$G(\mathbf{q_B}) = \ln \sum_{j \in \mathbf{B}} e^{-q_j}\pi_j(\mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}), \tag{5.24}$$

satisfying (5.19). This function fails to satisfy SS 5.6. We conclude that TPCS 5.7, or SS 5.6, are essential if the condition for a probabilistic choice system to be consistent with RUM is to be nonvacuous.

24. An example is the DOGIT model of Gaudry (1977), obtainable as a mixture of a multinomial logit and captive population PCS.

5.9 Criteria for Parametric Probabilistic Choice Systems

Assuming a population of random utility maximizers as a maintained hypothesis, the practical question is how to construct parametric PCS suitable for econometric and policy analysis. The criteria one may wish to impose on parametric PCS, beyond consistency with RUM, are (1) sufficient flexibility to capture patterns of substitution between alternatives and (2) a structure and parameterization facilitating estimation and computation. One approach to generating parametric PCS is to start from a parametric RUM model and derive the choice probabilities constructively. The primary drawback to this approach is that for many parametric RUM, the construction of the choice probabilities is analytically intractable, or results in functional forms that are impractical for empirical computation and analysis.

A second approach to specifying parametric PCS is to start from a practical system of choice probabilities and verify constructively or indirectly its consistency with RUM. One useful method is to test consistency with the sufficient conditions TPCS given in the preceding section.

5.10 Specification of Variables

Continuing the terminology of section 5.7, we consider an individual with a vector of measured individual characteristics $\mathbf{s}$, one component of which is income y. The individual faces alternatives characterized by vectors of measured nonprice attributes $\mathbf{w}_i$ and a budget constraint $q_i + \mathbf{r} \cdot \mathbf{x} = y$, where q_i and $\mathbf{r}$ are the prices of the discrete alternatives and divisible commodities, respectively. The individual has a utility function of $(\mathbf{x}, \mathbf{w}_i)$ which varies over the population. Without loss of generality, we can attribute this variation in utility to a vector of unmeasured nonprice attributes $\boldsymbol{\omega}_i$ and a vector of unmeasured individual characteristics $\boldsymbol{\sigma}$. Let $U(\mathbf{x}, \mathbf{w}_i, \mathbf{s}, \boldsymbol{\omega}_i, \boldsymbol{\sigma})$ denote this utility function.

The vector $\mathbf{w}_i$ contains information on real, or intrinsic, properties of the alternative, and in addition, nominal, or extrinsic, information such as labels attached by the observer for identification purposes. For example, a travel mode may be described by real variables such as time and number of transfers, as well as nominal labels such as bus, express, or alternative 4. It is reasonable to postulate that behavior depends solely on real proper-

ties of an alternative. However, an observed nominal label which is correlated with an unobserved real variable may appear in the population to be related to choice behavior. For example, a label that identifies a transportation mode as "bus" may be correlated with an unobserved real variable measuring the schedule flexibility of alternative modes and thus may act as a proxy for the unobserved real variable. The similarity of alternatives should also be perceived by individuals in real terms, but nominal classifications may act as proxies for the unmeasured real variables.

Some analyses of choice behavior, such as tests of the RUM hypothesis and the nature of similarity perceptions, can be carried out in PCS where alternatives are characterized partially or completely in nominal terms. This is particularly true in experimental applications where the universe of distinct alternatives is finite and saturated models are used where each alternative has a nominal label. However, analysis of economic behavior and forecasting is best done in real models. For example, knowledge of the historical effect of nominal variables, reflecting underlying unobserved real effects, is of little use in forecasting when unmeasured real variables change.

Empirical experience in travel demand forecasting (McFadden et al. 1977) suggests that it is difficult to construct purely real models using conventional market data alone. In terms of research strategy this suggests that most models using existing data will require nominal variables, but with their use limited, and that data collection should emphasize more extensive measurement of real variables.

5.11 Functional Form

The primary issues in choice of a functional form for a PCS are computational practicality and flexibility in representing patterns of similarity across alternatives. Practical experience suggests that functional forms that allow similar patterns of interalternative substitution will give comparable fits to existing economic data sets. Of course, laboratory experimentation or more comprehensive economic observations may make it possible to differentiate the fit of functional forms with respect to characteristics other than flexibility.

Currently three major families of concrete functional forms for PCS have been developed in the literature. These are logit models based on the work of Luce (1959), probit models based on the work of Thurstone (1927), and elimination models based on the work of Tversky (1972). Figure 5.1 outlines these families; the members are defined in sections 5.12 through 5.18.

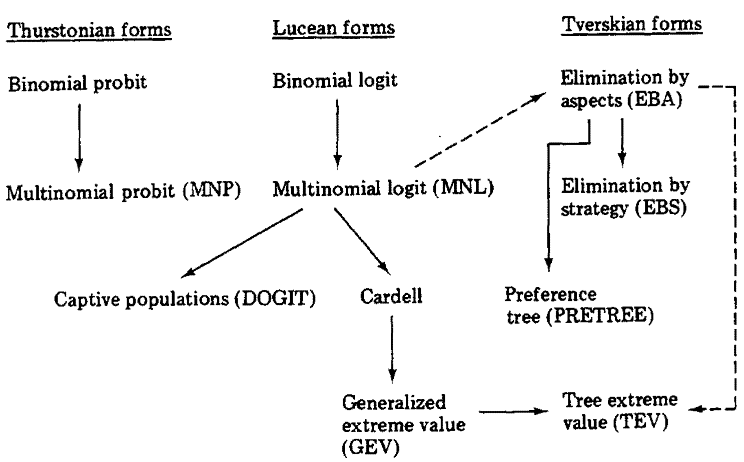


Figure 5.1
Concrete functional forms

In considering probit, logit, and related models, it is useful to quantify the random utility-maximizing hypothesis in the following terms: consider a choice set $\mathbf{B} = \{1, \ldots, m\} \in \mathscr{B}$. Alternative i has a column vector of observed attributes $\mathbf{z}_i$ and an associated utility $u_i = \boldsymbol{\alpha}'\mathbf{z}_i$, where $\boldsymbol{\alpha}$ is a vector of taste weights. Assume $\boldsymbol{\alpha}$ to have a parametric probability distribution with parameter vector $\boldsymbol{\theta}$, and let $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\theta})$ denote the mean and covariance matrix of $\boldsymbol{\alpha}$. Let $\mathbf{z}_\mathbf{B} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$ denote the array of observed attributes of the available alternatives. Then the vector of utilities $\mathbf{u}_\mathbf{B} = (u_1, \ldots, u_m)$ has a multivariate probability distribution with mean $\boldsymbol{\beta}'\mathbf{z}_\mathbf{B}$ and covariance matrix $\mathbf{z}'_\mathbf{B}\boldsymbol{\Omega}\mathbf{z}_\mathbf{B}$. The choice probability $P(i \mid \mathbf{B}, \mathbf{s})$ for alternative i, also written $P(i \mid \mathbf{z}_\mathbf{B}, \boldsymbol{\theta})$, then equals the probability of drawing a vector $\mathbf{u}_\mathbf{B}$ from this distribution such that $u_i \geqq u_j$ for $j \in \mathbf{B}$. For calculation note that $\mathbf{u}_{\mathbf{B}-i} = (u_1 - u_i, \ldots, u_{i-1} - u_i, u_{i+1} - u_i, \ldots, u_m - u_i)$ has a multivariate distribution with mean $\boldsymbol{\beta}'\mathbf{z}_{\mathbf{B}-i}$ and covariance matrix $\mathbf{z}'_{\mathbf{B}-i}\boldsymbol{\Omega}\mathbf{z}_{\mathbf{B}-i}$, where $\mathbf{z}_{\mathbf{B}-i} = (\mathbf{z}_1 - \mathbf{z}_i, \ldots, \mathbf{z}_{i-1} - \mathbf{z}_i, \mathbf{z}_{i+1} - \mathbf{z}_i, \ldots, \mathbf{z}_m - \mathbf{z}_i)$,

and that $P(i \mid \mathbf{z_B}, \boldsymbol{\theta})$ equals the nonpositive orthant probability for this $(m-1)$-dimensional distribution.[25]

The vector of attributes $\mathbf{z}_i$ of an alternative in this formulation is a function of the raw data $(\mathbf{q}_i, \mathbf{w}_i, \mathbf{r}, \mathbf{s})$, where $(q_i, \mathbf{w}_i)$ measure characteristics of the alternative and $(\mathbf{r}, \mathbf{s})$ characteristics of the individual and the background economic environment. Since any continuous (indirect) utility function can be approximated on a compact set to any desired degree of accuracy by an appropriate linear-in-parameters specification, and $\mathbf{z}_i$ can incorporate complex transformations and interactions of the raw data, there is virtually no loss of generality in assuming the utility structure $u_i = \boldsymbol{\alpha}'\mathbf{z}_i$.[26] Note that if $\mathbf{z}_i$ includes a (nominal) dummy variable specific to alternative i, then the corresponding component of $\boldsymbol{\alpha}$ can be interpreted as the contribution to utility of unobserved attributes of this alternative. The condition that u_i be homogeneous of degree zero in the prices $(q_i, \mathbf{r})$ will be met by requiring that $\mathbf{z}_i$ be homogeneous of degree zero

25. Suppose $\boldsymbol{\alpha}$ has a multivariate density $g(\boldsymbol{\alpha})$, with characteristic function

$$\gamma(\mathbf{t}) = \int_{-\infty}^{+\infty} e^{i\mathbf{t}'\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) d\boldsymbol{\alpha}.$$

Then $\mathbf{u}_{\mathbf{B}-j} = \boldsymbol{\alpha}'\mathbf{z}_{\mathbf{B}-j}$ has the characteristic function $Ee^{i\boldsymbol{\alpha}'\mathbf{z}_{\mathbf{B}-j}\boldsymbol{\tau}} = \gamma(\mathbf{z}_{\mathbf{B}-j}\boldsymbol{\tau})$, and associated density

$$h(\mathbf{u}_{\mathbf{B}-j}) = (2\pi)^{-m+1} \int_{-\infty}^{+\infty} e^{-i\boldsymbol{\tau}'\mathbf{u}_{\mathbf{B}-j}} \gamma(\mathbf{z}_{\mathbf{B}-j}\boldsymbol{\tau}) d\boldsymbol{\tau}$$

Then

$$P(j \mid \mathbf{z_B}, \boldsymbol{\theta}) = (2\pi)^{-m+1} \int_{\mathbf{u}_{\mathbf{B}-j}=-\infty}^{0} \int_{-\infty}^{+\infty} e^{-i\boldsymbol{\tau}'\mathbf{u}_{\mathbf{B}-j}} \gamma(\mathbf{z}_{\mathbf{B}-j}\boldsymbol{\tau}) d\boldsymbol{\tau}\, d\mathbf{u}_{\mathbf{B}-j}.$$

Starting from a density g with a closed form characteristic function, these formulas can be used in a few cases to obtain simple closed forms or expansions for the choice probabilities. More generally these formulas can be used, with suitable transformations of variables, to obtain numerical integrals of dimension $2(m-1)$ or less for the choice probabilities. However, some special structure is usually needed to make this approach practical.

26. Suppose indirect utility can be written in the form $U(\mathbf{t}_i, \boldsymbol{\tau}_i)$, with $\mathbf{t}_i = (q_i, \mathbf{w}_i, \mathbf{r}, \mathbf{s})$ and $\boldsymbol{\tau}_i$ a vector of unobserved attributes of alternatives and individual characteristics that

in $(q_i, \mathbf{r})$, and that the distribution of $\boldsymbol{\alpha}$ be invariant with respect to price level.

5.12 The Luce Model

A PCS that permits easy computation and interpretation but has a very restrictive pattern of interalternative substitution is the multinomial logit, MNL, form

$$P(i \mid \mathbf{z_B}, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}_i}}{\sum_{j \in \mathbf{B}} e^{\boldsymbol{\beta}'\mathbf{z}_j}}. \tag{5.25}$$

This form was first suggested by Luce (1959) as a psychological choice model derived from an axiom of independence from irrelevant alternatives, IIA:[27] If $i \in \mathbf{A} \subseteq \mathbf{B}$; then[28]

$$P(i \mid \mathbf{z_B}, \boldsymbol{\beta}) = P(i \mid \mathbf{z_A}, \boldsymbol{\beta}) P(\mathbf{A} \mid \mathbf{z_B}, \boldsymbol{\beta}). \tag{5.26}$$

determine the utility function. As noted earlier, variations in the utility functions of individuals can always be attributed to an unobserved vector $\tau_\mathbf{I}$. Suppose $\mathbf{T}$ and $\mathbf{T}^*$ are compact sets of $\mathbf{t}_i$ and τ_i, respectively. Suppose U is uniformly Lipschitzian in $\mathbf{t}_i$ on $\mathbf{T} \times \mathbf{T}^*$ with constant M; that is, $| U(t_i, \tau_i) - U(t_i', \tau_i) | \leqq M \mid t_i - t_i' \mid$ for $t_i, t_i' \in \mathbf{T}$ and $\tau_i \in \mathbf{T}^*$. Suppose $\mathbf{T}$ is defined—by translation, scaling, and extension if necessary—to equal $\mathbf{T} = \{\mathbf{t}_i \in \mathbf{R}^n \mid t_i \geqq 0, \Sigma_{j=1}^n t_{ij} \leqq 1\}$. Let $\mathbf{K}$ be the set of integer vectors $(k_1, \ldots, k_n)$ with $(k_1/N, \ldots, k_n/N) \in \mathbf{T}$, and define

$$z_{k_1 \cdots k_n}(\mathbf{t}_i) = t_{i1}^{k_1} \cdots t_{in}^{k_n} (1 - t_{i1} - \cdots - t_{in})^{N - k_1 - \cdots - k_n} N!/k_1! \ldots k_n!$$

Consider an approximation $\hat{U}$ to U defined by

$$\hat{U}(\mathbf{t}_i, \boldsymbol{\tau}_i) = \sum_{(k_1, \cdots, k_n) \in \mathbf{K}} \alpha_{k_1 \cdots k_n}(\boldsymbol{\tau}_i) z_{k_1 \cdots k_n}(\mathbf{t}_i),$$

with $\alpha_{k_1 \ldots k_n}(\tau_i) = U((k_1/N, \ldots, k_n/N), \tau_i)$. Given $\varepsilon > 0$, if $N \geqq nM^2/\varepsilon^2$, then $| U(\mathbf{t}_i, \tau_i) - \hat{U}(\mathbf{t}_i, \tau_i) | < \varepsilon$ on $\mathbf{T} \times \mathbf{T}^*$ (see McFadden 1978b, p. 236). Since $\mathbf{T}^*$ can be chosen so that the probability of unobserved variables falling outside $\mathbf{T}^*$ is as small as we please, the RUM $\hat{U}$ will yield a PCS which is as close as we please to the PCS generated by U. Note that this justification from approximation theory for a linear-in-parameters form does not imply that this approach is efficient, or even practical, for all applications.

27. The binary logistic curve is of much earlier vintage.

28. $P(\mathbf{A} \mid \mathbf{z_B}, \boldsymbol{\theta}) = \sum_{j \in \mathbf{A}} P(j \mid \mathbf{z_B}, \boldsymbol{\theta})$.

This system satisfies TPCS and can be derived from a social surplus function

$$G(\mathbf{q_B}) = \ln \sum_{j \in \mathbf{B}} e^{-q_j} \tag{5.27}$$

where $q_j = -\boldsymbol{\beta}'\mathbf{z}_j$.

The Luce model was shown by Marschak (1960) to be consistent with RUM. A constructive demonstration due to Marley is reported in Luce and Suppes (1965); see also McFadden (1973) and Yellot (1977). Specifically, if the u_i are independently distributed, with

$$\mathrm{Prob}\,[u_i \leqq \mu] = e^{-e^{-\mu-\boldsymbol{\beta}'\mathbf{z}_i}}, \tag{5.28}$$

then (5.2) yields (5.25) by an easy integration.[29] The distribution (5.28) is called a type I extreme value, or Weibull, distribution.

An implication of the MNL form is that all cross elasticities are equal; that is, for $i \neq j$,

$$\frac{\partial \ln P(i \mid \mathbf{z_B}, \boldsymbol{\theta})}{\partial \ln z_{jk}} = \beta_k z_{jk} P(j \mid \mathbf{z_B}, \boldsymbol{\theta}), \tag{5.29}$$

and the elasticity does not depend on i. This property is not plausible in some choice situations; see Debreu (1960) and McFadden, Tye, and Train (1977). The lack of flexibility of the Luce model is characteristic of a wider class of models satisfying the following property, called order independence: if $i, j \in \mathbf{A}$, $i, j \notin \mathbf{B}$, and $k \in \mathbf{B}$, then $P(i \mid \mathbf{A}) \geqq P(j \mid \mathbf{A})$ if and only if $P(k \mid \mathbf{B} \cup \{i\}) \leqq P(k \mid \mathbf{B} \cup \{j\})$.[30] A classic example shows that models satisfying order independence yield implausible conclusions when there are strong contrasts in the similarity of the alternatives. Suppose the alternatives are trips by auto (1), red bus (2), or blue bus (3). Suppose individuals treat the two buses as virtually equivalent and at prevailing travel times and costs divide evenly between auto and bus, so that $p_{12} = p_{13} = p_{23} = 0.5$, $p_{123} = 0.5$, and $p_{213} = p_{312} = 0.25$.[31] Consider

29. In the terminology of the preceding section with $u_i = \boldsymbol{\alpha}'\mathbf{z}_i$ and $\boldsymbol{\alpha}$ random, assume that the first m components of $\mathbf{z}$ are dummy variables for the m alternatives, that $\alpha_{m+1}, \ldots$ are nonrandom, and that $(\alpha_1, \ldots, \alpha_m)$ are independently distributed, with $E\alpha_i = \beta_i$ and $\mathrm{Prob}[\alpha_i - \beta_i \leqq \mu] = \exp[-e^{-\mu}]$. Note that at least one normalization, say, $\Sigma_{i=1}^m \beta_i = 0$, is needed for identification. In applications some or all of these β_i may be restricted to zero.

30. We assume with little loss of empirical generality that all choice probabilities are positive. This property was introduced by Tversky (1972a).

31. Define $p_{ij} = P(i \mid \{i, j\})$ and $p_{ijk} = P(i \mid \{i, j, k\})$.

$\mathbf{A} = \{1, 2, 3\}$ and $\mathbf{B} = \{3\}$. By order independence $p_{123} > p_{213}$ implies $p_{31} < p_{32}$, contradicting the probabilities we have constructed in light of the pattern of similarity of the alternatives.

Tversky (1972) has shown that order independence is equivalent to a property of PCS termed simple scalability: for $\mathbf{B} = \{1, \ldots, m\}$ the choice probabilities can be written as functions $P(i \mid \mathbf{B}, \mathbf{s}) = \pi_i(q_1(\mathbf{r}, \mathbf{w}_1, \mathbf{s}), \ldots, q_m(\mathbf{r}, \mathbf{w}_m, \mathbf{s}))$ of scale values (undesirability indices) $q_i(\mathbf{r}, \mathbf{w}_i, \mathbf{s})$ of the alternatives, with π_i increasing in q_j for $j \neq i$, and decreasing in q_i, π_i going to zero when $q_i \to +\infty$, and the form of the function π_i depending on the attributes of the alternatives solely through the scale values. Then clearly functional forms that are sufficiently flexible to accommodate cases of the red bus/blue bus type must depend on the orientation of alternatives in attribute space, as well as scalar measures of their desirability. For example, choice systems satisfying TPCS will be simply scalable if the social surplus function has the weakly separable form $G(h_1(q_1, \mathbf{w}_1, \mathbf{r}, \mathbf{s}), \ldots, h_m(q_m, \mathbf{w}_m, \mathbf{r}, \mathbf{s}), \mathbf{r}, \mathbf{s})$. Similarly PCS derived from RUM of the form described in the preceding section will be simply scalable when the attributes of alternatives shift only the mean of the distribution of utility levels.

Estimation of the multinomial logit model is discussed in McFadden (1973). A method for guaranteeing convergence of maximum likelihood estimation algorithms is discussed in section 5.20.

Because of its computational simplicity, the multinomial logit model has been a primary focus of attempts at functional generalizations. Some of these are discussed in section 5.15.

5.13 Thurstone's Model V

Thurstone (1927) proposed a random utility model for psychometric choice in which the utility levels of the alternatives are normally distributed. For binary choice this yields the probit model widely used in analysis of binary data; see Finney (1964) and Cox (1970). Multinomial probit, MNP, models are discussed in Bock and Jones (1968), McFadden (1976), Hausman and Wise (1976), Daganzo, Bouthelier, and Sheffi (1976), and Lerman and Manski, chapter 7.

Supose the utility of alternative i is $u_i = \boldsymbol{\alpha}'\mathbf{z}_i$, where $\boldsymbol{\alpha}$ is multivariate normal with mean $\boldsymbol{\beta}$ and covariance matrix $\mathbf{AA}'$. Additive normal variations in utility are incorporated in this formulation as random

coefficients of alternative specific dummy variables contained in $\mathbf{z}$. The vector $\mathbf{u}_{\mathbf{B}}$ is multivariate normal with mean $\mathbf{z}_{\mathbf{B}}\boldsymbol{\beta}$ and covariance matrix $\mathbf{z}_{\mathbf{B}}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}}'$. The choice probabilities satisfy

$$P(i \mid \mathbf{z}_{\mathbf{B}}, \boldsymbol{\beta}, \mathbf{A}) = N(-\mathbf{z}_{\mathbf{B}-i}\boldsymbol{\beta}, \mathbf{z}_{\mathbf{B}-i}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}-i}'), \tag{5.30}$$

where $N(\boldsymbol{\varepsilon}_{\mathbf{B}}, \boldsymbol{\Omega}_{\mathbf{B}})$ is the multivariate normal cumulative distribution function, with zero mean and covariance matrix $\boldsymbol{\Omega}_{\mathbf{B}}$, evaluated at $\boldsymbol{\varepsilon}_{\mathbf{B}}$, for any set of alternatives $\mathbf{B}$. This general structure and notation for the PCS follow from section 5.11 and the property that linear transformations of normal vectors are again normal.

In the case that alternative-specific dummies are included among the attributes, it is convenient to redefine $u_i = -q_i + \boldsymbol{\alpha}'\mathbf{z}_i$, where $-q_i$ is the mean of the alternative-specific effect. Then

$$P(i \mid \mathbf{z}_{\mathbf{B}}, \mathbf{q}_{\mathbf{B}}, \boldsymbol{\beta}, \mathbf{A}) = N(\mathbf{q}_{\mathbf{B}-i} - \mathbf{z}_{\mathbf{B}-i}\boldsymbol{\beta}, \mathbf{z}_{\mathbf{B}-i}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}-i}'), \tag{5.31}$$

and the choice probabilities satisfy TPCS, with a social surplus function [32]

$$\begin{aligned} G(\mathbf{q}_{\mathbf{B}}, \mathbf{z}_{\mathbf{B}}\boldsymbol{\beta}, \mathbf{z}_{\mathbf{B}}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}}') &= \int_{t=-\infty}^{+\infty} [N(-\mathbf{z}_{\mathbf{B}}\boldsymbol{\beta} + t, \mathbf{z}_{\mathbf{B}}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}}') \\ &\qquad - N(\mathbf{q}_{\mathbf{B}} - \mathbf{z}_{\mathbf{B}}\boldsymbol{\beta} + t, \mathbf{z}_{\mathbf{B}}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}}')]\,dt \\ &= -q_1 - \int_{t=0}^{\infty} [N(\mathbf{q}_{\mathbf{B}-1} - \mathbf{z}_{\mathbf{B}-1}\boldsymbol{\beta} + t, \mathbf{z}_{\mathbf{B}-1}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}-1}') \\ &\qquad - N(-\mathbf{z}_{\mathbf{B}-1}\boldsymbol{\beta} + t, \mathbf{z}_{\mathbf{B}-1}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}-1}')]\,dt. \end{aligned} \tag{5.32}$$

Evaluation of the choice probabilities or the social surplus function requires numerical integration or approximation of $(m-1)$-dimensional multivariate normal orthant probabilities. This is practical with conventional expansions for $m \leqq 3$ but generally impractical for $m > 5$; see Hausman and Wise (1978). A procedure due to Clark (1961) that

32. Recall that in the notation adopted beginning in section 5.11 the attributes $(\mathbf{z}_i, \mathbf{q}_i)$ are assumed to be homogeneous of degree zero in economic prices. Thus, if q_i is the price of the discrete alternative, it is here assumed to be a relative price. The conditions SS and TPCS in section 2.6 are stated in terms of a vector of absolute prices and impose the restrictions that the social surplus function and choice probabilities be homogeneous of degree one and zero, respectively, in these prices. For the current application the prices in SS and TPCS should be reinterpreted as being relative, and the homogeneity conditions in SS 5.2 and TPCS 5.2 should be ignored.

approximates the maximum of two normal variates by a normal variate is reasonably accurate for $m \leqq 10$ when the underlying variates have comparable variances and nonnegative correlations; see Daganzo, Bouthelier, and Sheffi (1977), Lerman and Manski, chapter 7, and Horowitz (1979).

When the MNP model has the form (5.31) with only the coefficients of alternative-specific dummy variables distributed randomly, the covariance matrix $\mathbf{z}_{\mathbf{B}-i}\mathbf{A}\mathbf{A}'\mathbf{z}_{\mathbf{B}-i}'$ depends only on the nominal labels of alternatives contained in $\mathbf{z}$, by PCS 5.2. If in this case there is no plausible link between nominal labels and degree of similarity, the covariance matrix will have a structure independent of the alternatives in the choice set, making the model simply scalable, and hence subject to the criticisms given the Luce model. More generally, with stochastic variation in coefficients of $\boldsymbol{\alpha}$ other than dummy coefficients, the MNP model permits quite flexible patterns of substitution across alternatives and components of variance structure for taste variations. The primary difficulty in application of the MNP model is the lack of practical, accurate methods for approximating the choice probabilities when the number of alternatives is large. There are also some technical issues in formulating iterative algorithms for maximum likelihood estimation in MNP models; see section 5.20.

5.14 Tversky Elimination Models

Choice can be viewed as a process in which alternatives are eliminated from the choice set until a single alternative remains. Formally, an elimination model is defined by a transition probability $Q: \mathscr{B} \times \mathscr{B} \times \mathbf{S} \times \mathbf{T} \to [0, 1]$, with $\mathbf{T} = \{1, 2, \ldots\}$ and with $Q(\mathbf{A} \mid \mathbf{B}, \mathbf{s}, t)$ interpreted as the probability that, starting from a set of alternatives $\mathbf{B}$ at step t, decision makers will reach in the next step the set $\mathbf{A}$ by eliminating some alternatives. The transition probabilities must satisfy $Q(\mathbf{A} \mid \mathbf{B}, \mathbf{s}, \mathbf{t}) \geqq 0$, $Q(\varnothing \mid \mathbf{B}, \mathbf{s}, t) = 0$, $Q(\mathbf{B} \mid \mathbf{B}, \mathbf{s}, t) < 1$, and $\Sigma_{\mathbf{A} \in \mathscr{B}}^{\mathbf{A} \subseteq \mathbf{B}} Q(\mathbf{A} \mid \mathbf{B}, \mathbf{s}, t) = 1$. Choice probabilities equal the sum, over all possible chains, of transition probabilities for the chain. When the transition probabilities are stationary (independent of t), the choice probabilities are given by the recursion formula

$$P(i \mid \mathbf{B}, \mathbf{s}) = \sum_{\mathbf{A} \in \mathscr{F}_{\mathbf{B}}^{*}} Q(\mathbf{A} \mid \mathbf{B}, \mathbf{s}) P(i \mid \mathbf{A}, \mathbf{s}), \tag{5.33}$$

where $\mathscr{F}_{\mathbf{B}}^{*} = \{\mathbf{A} \in \mathscr{B} \mid \mathbf{A} \subseteq \mathbf{B}, \emptyset \neq \mathbf{A}\}$.

Concrete elimination models are specified by parameterizing the transition probabilities Q. Tversky (1972a, 1972b) has introduced an important class of elimination by aspect, EBA, models in which there is associated with each $\mathbf{A} \in \mathscr{B}$ a nonnegative scale value $v_{\mathbf{A}} = V(\mathbf{z}_{\mathbf{A}})$, and the transition probabilities satisfy

$$Q(\mathbf{A} \mid \mathbf{B}, \mathbf{s}) = \frac{v_{\mathbf{A}}}{\sum_{\mathbf{C} \in \mathscr{F}^{*}_{\mathbf{B}}} v_{\mathbf{C}}} \tag{5.34}$$

for $\mathbf{A} \in \mathscr{F}^{*}_{\mathbf{B}}$. Tversky shows that this model can be interpreted as the result of a choice process in which decision makers sample from some distribution over aspects of alternatives, eliminating alternatives that fail to have the sampled aspect. The scale value $v_{\mathbf{A}}$ is interpreted as the probability of drawing an aspect that is unique to $\mathbf{A}$ and common within $\mathbf{A}$. The EBA model is consistent with a population of preference maximizers, each with lexicographic preferences over aspects.

The Luce model is a special case of the EBA family, obtained when $v_{\mathbf{A}} = 0$ for $\mathbf{A}$ containing more than one element. More generally the EBA model can accommodate complex patterns of substitutability of alternatives, with $v_{\mathbf{A}}$ measuring the similarity of the alternatives in $\mathbf{A}$. An even more general family of elimination models that are consistent with random preference maximization and permit nonstationary transition probabilities can be defined by considering strategies for selecting aspects. These elimination-by-strategy, EBS, models are discussed in section 5.24.

The EBA functional form has considerable potential for econometric applications. When the scale functions V are log-linear in parameters, $\ln V(\mathbf{z}_{\mathbf{A}}) = \boldsymbol{\beta}'_{\mathbf{A}}\mathbf{z}_{\mathbf{A}}$, the choice probabilities can be written as sums of products of MNL forms. Maximum likelihood estimation could be carried out for such systems with relatively minor modifications of current MNL computer programs. One drawback of EBA for econometric applications is that the motivation for the model provides little guidance for parametric specification of the scale function V.

5.15 Generalized Extreme Value Models

A number of authors have sought variants of the MNL form that retain its computational advantages while permitting more flexible patterns of substitution. Most of these variants must be rejected because they are

inconsistent with RUM or fail to accommodate substitution patterns of the red bus/blue bus type.[33]

The MNL model can be derived from a RUM with independently extreme value-distributed utility levels. One approach to generalizing the MNL model is to start from a more general multivariate extreme value distribution. Suppose alternative i has a scale value $q_i = -\boldsymbol{\beta}'\mathbf{z}_i$. The following result gives a large class of logitlike PCS. These are termed generalized extreme value, GEV, models:

THEOREM 5.2: For $\mathbf{B} = \{1, \ldots, m\} \in \mathscr{B}$, consider $H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s})$ satisfying

GEV 5.1: $H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s})$ is a nonnegative, linear homogeneous function of $\mathbf{y_B} \geqq \mathbf{0}$.

GEV 5.2: $\lim\limits_{y_i \to +\infty} H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s}) = +\infty$.

GEV 5.3: The mixed partial derivatives of H exist and are continuous, with nonpositive even and nonnegative odd mixed partial derivatives.

GEV 5.4: If $\mathbf{B} = \{i_1, \ldots, i_m\} \in \mathscr{B}$ and $\mathbf{B}' = \{i_1', \ldots, i_m', \ldots, i_{m+n}'\} \in \mathscr{B}$ satisfies $\mathbf{z}_{i_k} = \mathbf{z}_{i_k'}$ for $k = 1, \ldots, m$, then

$$H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s}) = H((\mathbf{y_B}, 0, \ldots, 0), \mathbf{z_{B'}}, \mathbf{s}).$$

If H satisfies GEV and $\mathbf{B} = \{1, \ldots, m\}$, then

$$F(\boldsymbol{\varepsilon}_\mathbf{B}, \mathbf{z_B}, \mathbf{s}) = \exp\{-H(e^{-\varepsilon_1}, \ldots, e^{-\varepsilon_m}, \mathbf{z_B}, \mathbf{s})\} \tag{5.35}$$

is a multivariate extreme value distribution. A random utility-maximizing model in which utility levels are distributed $F(\mathbf{u_B} + \mathbf{q_B}, \mathbf{z_B}, \mathbf{s})$ for $\mathbf{B} \in \mathscr{B}$ satisfies AIRUM and has a social surplus function

$$\begin{aligned} G(\mathbf{q_B}, \mathbf{z_B}, \mathbf{s}) &= E \max_{i \in \mathbf{B}} u_i - \gamma \\ &= \ln H(e^{-q_1}, \ldots, e^{-q_m}, \mathbf{z_B}, \mathbf{s}), \end{aligned} \tag{5.36}$$

where γ is a constant independent of $\mathbf{B}$, and choice probabilities satisfying

33. Models that may fail to be consistent with RUM are the cascade and maximum models of McFadden (1974), the universal logit model of McFadden (1975), and the fully competitive model of McLynn (1973). Models with insufficient flexibility to accommodate reasonable patterns of substitution are the cascade and fully competitive models and the DOGIT model of Gaudry (1977).

$$\pi_i(\mathbf{q_B}, \mathbf{z_B}, \mathbf{s}) = -\frac{\partial}{\partial q_i}\ln H(e^{-q_1}, \ldots, e^{-q_m}, \mathbf{z_B}, \mathbf{s}). \tag{5.37}$$

This theorem is proved by first showing that F in (5.35) is a cumulative probability distribution; that is, the range of F is the unit interval and the density f of F is nonnegative. The range condition follows from GEV 5.1 and GEV 5.2, and the nonnegativity of f can be established by induction on the order of a mixed partial derivative of F, using GEV 5.3. The formulae (5.36) and (5.37) for the social surplus function and for the choice probabilities follow by direct integration from (5.35). A formal proof is given in McFadden (1978a).

Property GEV 5.1 of H implies that the function G defined in (5.35) satisfies SS 5.1 and SS 5.3. Property SS 5.6 is a consequence of GEV 5.4. An induction argument using GEV 5.3 establishes SS 5.4. To show SS 5.5, note that

$$G_i = \frac{-H_i(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s})}{H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s})} \tag{5.38}$$

with $y_j = e^{-q_j+q_i}$. As $q_i \to -\infty$, $\mathbf{y_B}$ converges to a vector $\mathbf{y'_B}$ with one in component i, zeroes elsewhere. Since $H(\mathbf{y'_B}, \mathbf{z_B}, \mathbf{s}) = H_i(\mathbf{y'_B}, \mathbf{z_B}, \mathbf{s}) > 0$ by GEV 5.2, (5.38) has the limit -1 as $q_i \to -\infty$. Under the assumptions of this section the q_i are homogeneous of degree zero in absolute prices and SS 5.2 need not be imposed. Then G is a social surplus function, and theorem 5.1 gives an alternative, nonconstructive proof that an AIRUM model exists with PCS satisfying (5.37).

When the social surplus function (5.36) depends on the $\mathbf{z}_i$ only through the terms $\boldsymbol{\beta}'\mathbf{z}_i = -q_i$, the PCS is simply scalable. Hence GEV models with flexible crossalternative substitution require dependence of $H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s})$ on the orientation of alternatives $\mathbf{z_B}$.

The GEV model reduces to the Luce model when

$$H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s}) = \left[\sum_{i\in \mathbf{B}} y_i^{1/1-\sigma}\right]^{1-\sigma},$$

$0 \leqq \sigma < 1$. An example of a more general function satisfying GEV is

$$H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s}) = \sum_{\mathbf{C}\in \mathscr{F}^{*}_{\mathbf{B}}} a(\mathbf{z_C})\left[\sum_{i\in \mathbf{C}} y_i^{1/1-\sigma(\mathbf{z_C})}\right]^{1-\sigma(\mathbf{z_C})}, \tag{5.39}$$

with $a(\mathbf{z}_\mathbf{C}) \geqq 0$ and $0 \leqq \sigma(\mathbf{z}_\mathbf{C}) < 1$.[34] The PCS for (5.39) can be written

$$P(i \mid \mathbf{B}, \mathbf{s}) = \sum_{\mathbf{C} \in \mathscr{F}^{\bullet}_{\mathbf{B}}} P(i \mid \mathbf{C}, \mathbf{s}) Q(\mathbf{C} \mid \mathbf{B}, \mathbf{s}), \tag{5.40}$$

where

$$P(i \mid \mathbf{C}, \mathbf{s}) = \frac{e^{\boldsymbol{\beta}'\mathbf{z}_i/1 - \sigma(\mathbf{z}_\mathbf{C})}}{\sum_{j \in \mathbf{C}} e^{\boldsymbol{\beta}'\mathbf{z}_j/1 - \sigma(\mathbf{z}_\mathbf{C})}} \quad \text{for } i \in \mathbf{C}, \tag{5.41}$$

and

$$Q(\mathbf{C} \mid \mathbf{B}, \mathbf{s}) = \frac{a(\mathbf{z}_\mathbf{C}) e^{(1 - \sigma(\mathbf{z}_\mathbf{C}))h_\mathbf{C}}}{\sum_{\mathbf{A} \in \mathscr{F}^{\bullet}_{\mathbf{B}}} a(\mathbf{z}_\mathbf{A}) e^{(1 - \sigma(\mathbf{z}_\mathbf{A}))h_\mathbf{A}}}, \tag{5.42}$$

with

$$h_\mathbf{C} = \ln \sum_{j \in \mathbf{C}} e^{\boldsymbol{\beta}'\mathbf{z}_j/1 - \sigma(\mathbf{z}_\mathbf{C})}. \tag{5.43}$$

This can be interpreted as a PCS in which decision makers invoke a subset of alternatives **C** from **B** and then select an alternative from **C**. Conditional choice of an alternative from a set **C** has choice probabilities (5.41) of the MNL form. Associated with a set **C** is an inclusive value $h_\mathbf{C}$ which equals the social surplus obtained from the MNL form (5.41). Choice probabilities for the invoked set **C** are multinomial logit functions of the inclusive values. The function $\sigma(\mathbf{z}_\mathbf{C})$ is a measure of the similarity of alternatives within **C**. When the alternatives in **C** are very similar and $\sigma(\mathbf{z}_\mathbf{C})$ is near one, the conditional choice probability (5.41) selects with high probability the alternative with the highest value in **C** of $\boldsymbol{\beta}'\mathbf{z}_i$. Then $h_\mathbf{C}$ is approximately $\max_{i \in \mathbf{C}} \boldsymbol{\beta}'\mathbf{z}_i/(1 - \sigma(\mathbf{z}_\mathbf{C}))$, and in the choice of an invoked set using the probabilities (5.42), the set **C** is assessed approximately as if it contained a single alternative with a scale value $\max_{i \in \mathbf{C}} \boldsymbol{\beta}'\mathbf{z}_i$.

Functions of the form (5.39) can also be nested to multiple levels to yield a broader class of functions. For example, the two-level nested function

34. GEV 5.2 requires for each $i \in \mathbf{B}$ that $a(\mathbf{z}_\mathbf{C}) > 0$ for some subset **C** of **B** containing i. A similar condition is required on (5.44).

$$H(\mathbf{y_B}, \mathbf{z_B}, \mathbf{s}) = \sum_{\phi \neq \mathbf{C} \subseteq \mathbf{B}} a_1(\mathbf{z_C}) \left[\sum_{\phi \neq \mathbf{D} \subseteq \mathbf{C}} a_2(\mathbf{z_D}) \left[\sum_{i \in \mathbf{D}} y_i^{1/1-\sigma_2(\mathbf{z_D})} \right]^{(1-\sigma_2(\mathbf{z_D}))/1-\sigma_1(\mathbf{z_C})} \right]^{1-\sigma_1(\mathbf{z_C})} \quad (5.44)$$

satisfies GEV when $a_1(\mathbf{z_C}) \geqq 0$, $a_2(\mathbf{z_D}) \geqq 0$, and $0 \leqq \sigma_2(\mathbf{z_D}) \leqq \sigma_1(\mathbf{z_C}) < 1$ for $\mathbf{D} \subseteq \mathbf{C}$. The PCS generated by (5.44) can be written, analogously to (5.40) through (5.43), as a sum of products of transition probabilities, with each transition probability corresponding to a level in the nest and expressable in an MNL form giving the choice at the next level of the nest, conditioned on the invoked set at this level. The variables in this transition probability are inclusive values of the next stage alternatives. The bounds on the parameters in (5.44) imply that the coefficients of inclusive value at each level lie in the unit interval; this condition can also be shown using theorem 5.1 to be necessary for consistency of the PCS, from (5.37) and (5.44), with AIRUM.

Choice probabilities of the form (5.40) through (5.43) were apparently first derived for a special case by Cardell (1977). Alternative derivations of PCS with the nested MNL structure (5.40) through (5.43) have been given independently by Ben-Akiva and Lerman (1977) and Daly and Zachary (1976).

As the example (5.40) through (5.43) and its obvious extensions make clear, GEV models can be interpreted as elimination models with nonstationary transition probabilities, or elimination-by-strategy models in the terminology of section 5.24. The transition probabilities have MNL forms as in the elimination-by-aspect model (5.34) but with a different definition of the scale values. Inspection suggests that EBA and GEV models are roughly comparable in flexibility and complexity.

5.16 Preference Trees

The elimination-by-aspects and generalized extreme value models both permit very general patterns of similarities between alternatives. Psychological analysis suggests that judgments of similarity often exhibit a more restricted structure in which statements such as "A is more like B than it is like C" have behavioral meaning; see particularly Sattath and Tversky

(1977), and also Rumelhart and Greeno (1971), Edgell et al. (1973), and Sjöberg (1975). With this structure alternatives can be arrayed in a preference tree with the least similar alternatives on the most distant branches; an example is given in figure 5.2. In economic choice it is reasonable to postulate that crossalternative substitutability is related to perceived similarity, so that alternatives on distant branches are the least substitutable. This suggests that choice in a tree be modeled as a process of transitions through a fixed hierarchy of nodes, eliminating undesirable branches until a single alternative is reached.[35] It is clear from figure 5.2 that preference trees can accommodate examples of red bus/blue bus type by making the bus alternatives very similar to each other.

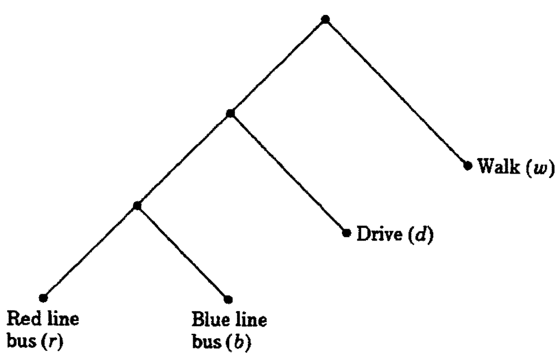


Figure 5.2
A preference tree for transportation modes

The general notation used in sections 5.14 and 5.15 to describe transitions specializes readily to preference trees and will be used for the model definition. (Later, for convenience in applications, we introduce an index notation for nodes.) A node in a preference tree is identified by the set of elemental alternatives at the ends of the branch below this node. This characterization can be applied to the tree formed by pruning away all branches not containing an alternative in a specified set of elemental alternatives.

35. When data are observed only on final choices, as is usual in economic applications, it is impossible to determine whether this elimination heuristic describes decision processes used by individuals. There is some evidence from verbal protocols in laboratory experiments supporting hierarchical elimination as a description of the cognitive choice process.

Any set $\mathbf{B} \in \mathscr{B}$ can be identified as the stem, or uppermost node, of the tree formed by pruning all branches not containing elements of $\mathbf{B}$. Define $\mathscr{F}_{\mathbf{B}}$ to be the family of nodes immediately below the stem of the pruned tree and $\mathscr{F}^*_{\mathbf{B}}$ to be the family of all nodes in the branch starting from $\mathbf{B}$. In figure 5.2, for example, $\mathscr{F}_{\{r,b,d,w\}} = \{\{w\},\{r,b,d\}\}$ and $\mathscr{F}^*_{\{r,b,d,w\}} = \{\{w\},\{r,b,d\}, \{d\}, \{r, b\}, \{r\}, \{b\}, \{r, b, d, w\}\}$, while $\mathscr{F}_{\{r,b,w\}} = \{\{w\}, \{r, b\}\}$ and $\mathscr{F}^*_{\{r,b,w\}} = \{\{w\}, \{r, b\}, \{r\}, \{b\}, \{r, b, w\}\}$.

We shall define a hierarchical elimination system, HES, for a specified preference tree to be one in which the PCS satisfies a recursion relation

$$P(i \mid \mathbf{B},\mathbf{s}) = \sum_{\mathbf{C} \in \mathscr{F}_{\mathbf{B}}} P(i \mid \mathbf{C},\mathbf{s}) Q(\mathbf{C} \mid \mathbf{B},\mathbf{s}) \tag{5.45}$$

for some family of transition probabilities betwen adjacent nodes. For each $\mathbf{B} \in \mathscr{B}$ and $i \in \mathbf{B}$, the preference tree identifies a unique sequence of nodes $\{i\} \equiv \mathbf{B}_0 \subseteq \mathbf{B}_1 \subseteq \cdots \subseteq \mathbf{B}_J \equiv \mathbf{B}$ with $\mathbf{B}_{j-1} \in \mathscr{F}_{\mathbf{B}_j}$. The choice probability can be written

$$P(i \mid \mathbf{B},\mathbf{s}) = Q(\mathbf{B}_0 \mid \mathbf{B}_1,\mathbf{s}) Q(\mathbf{B}_1 \mid \mathbf{B}_2,\mathbf{s}) \ldots Q(\mathbf{B}_{J-1} \mid \mathbf{B}_J,\mathbf{s}). \tag{5.46}$$

Then observations on choices are equivalent to observations on transitions, and maximum likelihood estimation of the parameters of an HES can be interpreted as maximum likelihood estimation of the parameters of the transition probabilities using observations on transitions. If the transition probabilities have computationally practical functional forms, fully efficient (full information) maximum likelihood estimation of the system parameters may be feasible. More generally it may be possible to formulate a sequence of computationally simple conditional maximum likelihood procedures, corresponding to levels in the tree, which yield consistent, but not in general efficient, estimators. These observations will apply to the two parametric specifications of HES based on the EBA and GEV models.

First consider the EBA model applied to preference trees. This specialization has been developed by Tversky and Sattath (1979), who derive its properties and report experimental evidence on its validity. Let $v_{\mathbf{C}} = V(\mathbf{z}_{\mathbf{C}})$ be a nonnegative scale value associated with a node $\mathbf{C}$, where $\mathbf{z}_{\mathbf{C}}$ is a vector of observed attributes of the alternatives in $\mathbf{C}$.[36] One can

36. The vector $\mathbf{z}$ is defined as in section 5.12 to incorporate the effects of attributes of alternatives, individual characteristics, and their interactions. We interpret $\mathbf{z}_{\mathbf{C}}$ as the vector of attributes common to or representative of the alternatives in $\mathbf{C}$.

interpret $v_{\mathbf{C}}$ as a measure of the set of aspects of alternatives that are unique to **C** and common within **C**. Define a nonnegative scale value $v^*_{\mathbf{C}}$ associated with the branch with stem **C**,

$$v^*_{\mathbf{C}} = \sum_{\mathbf{A} \in \mathscr{F}^*_{\mathbf{C}}} v_{\mathbf{A}}. \tag{5.47}$$

For $\mathbf{C} \in \mathscr{F}_{\mathbf{B}}$ define the transition probabilities

$$Q(\mathbf{C} \mid \mathbf{B}, \mathbf{s}) = \frac{v^*_{\mathbf{C}}}{\sum_{\mathbf{A} \in \mathscr{F}_{\mathbf{B}}} v^*_{\mathbf{A}}}. \tag{5.48}$$

We term a PCS satisfying (5.45) through (5.48) a hierarchical elimination-by-aspects, HEBA, model.

The EBA model (5.33) through (5.34) applied to a preference tree, with the specified scale values $v_{\mathbf{C}}$ for nodes in the tree and zero scale values for other subsets of the choice set, permits direct transitions from a node to any node in the branch below it, in contrast to the hierarchical protocol employed in HEBA. Despite this apparent difference the two models yield the same PCS. This can be seen for the example in figure 5.2. by writing out the choice probabilities. A general equivalence theorem has been proved by Tversky and Sattath. Since EBA is consistent with random preference maximization, this theorem estabishes that HEBA is also consistent. These authors refer to HEBA in either of its equivalent forms as a PRETREE model.

For econometric applications the scale functions V can be assumed with little loss of generality to be log-linear in parameters, ln $V(\mathbf{z}_{\mathbf{C}}) = \boldsymbol{\beta}'_{\mathbf{C}}\mathbf{z}_{\mathbf{C}}$ for nodes **C** in the tree. Then, as in the general EBA model, the transition probabilities can be written as sums of MNL functional forms, with $Q(\mathbf{B}_0 \mid \mathbf{B}_1, \mathbf{s})$ a simple MNL form, and the transition probability at a node $\mathbf{B}_j$ at level j in the tree depending on $\boldsymbol{\beta}_{\mathbf{B}_j}$ and on terms appearing in the transition probabilities at levels $1, \ldots, j-1$. Then a sequential procedure that will yield consistent parameter estimates under normal regularity conditions is to first estimate the parameters of $\boldsymbol{\beta}_{\mathbf{B}_0}$ by conditional maximum likelihood estimation applied to level 1 transitions, then estimate the parameters of $\boldsymbol{\beta}_{\mathbf{B}_1}$ using level 2 transitions, substituting the estimate of $\boldsymbol{\beta}_{\mathbf{B}_0}$ obtained at level 1, and so on. Computational formulae for this procedure can be developed analogously to the formulae given in section 5.22 for the following model.

Next consider the GEV model applied to preference trees. Define scale values $v_i = e^{-q_i} \equiv e^{\boldsymbol{\beta}'\mathbf{z}_i}$, where we assume with little loss of generality that v_i is log-linear in parameters $\boldsymbol{\beta}$ and use the notation $q_i = -\boldsymbol{\beta}'\mathbf{z}_i$. Define a function $\sigma(\mathbf{z}_\mathbf{C})$ measuring the similarity of alternatives at node $\mathbf{C}$. Assume $0 \leqq \sigma(\mathbf{z}_\mathbf{C}) < 1$, with increasing σ denoting greater similarity, or correlation of attributes. Let $\boldsymbol{\theta}_\mathbf{C} = 1 - \sigma(\mathbf{z}_\mathbf{C})$. Assign scale values to nodes using the recursion

$$v_\mathbf{A} = \left[\sum_{\mathbf{C} \in \mathscr{F}_\mathbf{A}} v_\mathbf{C}^{1/\theta_\mathbf{A}} \right]^{\theta_\mathbf{A}}. \tag{5.49}$$

The probabilistic choice system for this model then satisfies the recursion (5.45) for hierarchical elimination, with transition probabilities

$$Q(\mathbf{C} \mid \mathbf{B}, \mathbf{s}) = \frac{v_\mathbf{C}^{1/\theta_\mathbf{B}}}{\sum_{\mathbf{A} \in \mathscr{F}_\mathbf{B}} v_\mathbf{A}^{1/\theta_\mathbf{B}}}. \tag{5.50}$$

The conditions GEV require $\theta_\mathbf{A} \leqq \theta_\mathbf{B}$ for $\mathbf{A} \in \mathscr{F}_\mathbf{B}$. We term (5.45), (5.49), and (5.50) the tree extreme value, TEV, model. For $\mathbf{B} = \{1, \ldots, m\} \in \mathscr{B}$, this model has a social surplus function

$$G(\mathbf{q}_\mathbf{B}, \mathbf{z}_\mathbf{B}, \mathbf{B}) = \ln H(e^{-q_1}, \ldots, e^{-q_m}, \mathbf{z}_\mathbf{B}, \mathbf{B}), \tag{5.51}$$

where H is defined recursively by $H(y_i, \mathbf{z}_i, \{i\}) = y_i$ and for any node $\mathbf{A}$,

$$H(\mathbf{y}_\mathbf{A}, \mathbf{z}_\mathbf{A}, \mathbf{A}) = \left[\sum_{\mathbf{C} \in \mathscr{F}_\mathbf{A}} H(\mathbf{y}_\mathbf{C}, \mathbf{z}_\mathbf{C}, \mathbf{C})^{1/\theta_\mathbf{A}} \right]^{\theta_\mathbf{A}}. \tag{5.52}$$

Estimation of the TEV model is discussed in sections 5.17 through 5.18.

Since the HEBA and TEV models have similar structures and possible levels of parameterization, one would expect them to give similar fits to data. To test this conjecture numerically, we considered the simple three-alternative preference tree illustrated in figure 5.3. The HEBA and TEV models are fitted to binary choice probabilities that come from this tree; HEBA and TEV are each just identified by the binary probabilities. Then the multinomial choice probabilities implied by the models are compared. A simple MNP model is included in the comparison.

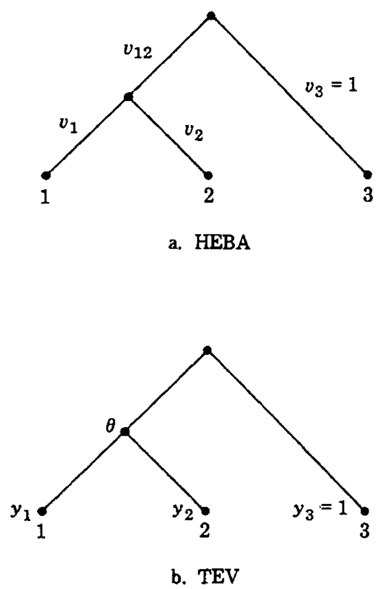


Figure 5.3
Simple HEBA and TEV preference trees

Let p_{ij} denote the binary choice probability of i over j, and P_i the multinomial choice probability. Assume the alternatives are numbered so that $p_{12} > 0.5$. Tversky and Sattath show that the preference tree configuration in figure 5.3 then implies a trinary condition,

$$\frac{p_{12}}{p_{21}} > \frac{\dfrac{p_{13}}{p_{31}}}{\dfrac{p_{23}}{p_{32}}} > 1. \tag{5.53}$$

The TEV model for this preference tree must also satisfy (5.53). Hence we compare these models for the set of binary choice probabilities satisfying (5.53).

The HEBA model for figure 5.3 can be written

$$p_{12} = \frac{v_1}{v_1 + v_2},$$

$$p_{13} = \frac{v_1 + v_{12}}{v_1 + v_3 + v_{12}},$$

$$p_{23} = \frac{v_2 + v_{12}}{v_1 + v_3 + v_{12}},$$

$$P_1 = \frac{v_1 + v_{12} p_{12}}{v_1 + v_2 + v_3 + v_{12}},$$

$$P_2 = \frac{v_2 + v_{12} p_{21}}{v_1 + v_2 + v_3 + v_{12}}, \tag{5.54}$$

where v_1, v_2, v_3, v_{12} are treated as parameters, with the normalization $v_3 = 1$. The TEV model is generated by the function

$$H(y_1, y_2, y_3) = (y_1^{1/\theta} + y_2^{1/\theta})^{\theta} + y_3, \tag{5.55}$$

with $p_{12} = \partial \ln H(y_1, y_2, 0)/\partial \ln y_1$, $P_1 = \partial \ln H(y_1, y_2, y_3)/\partial \ln y_1$, and so on. The parameters of the model are y_1, y_2, y_3, and θ, with the normalization $y_3 = 1$.

To form an MNP model with a similarity structure mimicing figure 5.3, we assume the multivariate normal random utility vector (u_1, u_2, u_3) has u_3 independent of u_1 and u_2. Imposing the trinary condition (5.53) implies a common variance for u_1 and u_2. Then this model also has three independent parameters.

Table 5.1 compares the multinomial probabilities from these three models for a selection of values of the binary probabilities. For the MNP model both the exact probabilities and approximate values obtained by the Clark method are given. Appendix 5.21 gives computational formulas.

The most striking feature of table 5.1 is the closeness of the multinomial probabilities predicted by HEBA, TEV, and MNP. The absolute deviation of HEBA and TEV for these cases is at most 0.0074, and the maximum relative deviation is 6 percent. The absolute deviation of TEV and MNP is at most 0.016. The relative deviation of TEV and MNP can rise to 22 percent for small probabilities but for probabilities over 0.1 is under 6 percent. We conclude that at least for simple preference trees such as figure 5.3, these models are for all practical purposes indistinguishable. Cases 4, 5, and 6 parallel the red bus/blue bus example, with case 4 corresponding to high similarity of the bus alternatives and case 6 to low similarity. All three models generate the intuitively plausible multinomial probabilities for these cases.

The Clark approximation to the MNP probabilities is quite inaccurate in a few cases, with absolute deviations as high as 0.1 and relative deviations

Table 5.1
A comparison of HEBA, TEV, and MNP choice probabilities for a simple preference tree

	Case			Nonsimilarity		P_1				P_2			
	p_{12}	p_{13}	p_{23}	θ	$1-\rho$	HEBA	TEV	MNP	Clark	HEBA	TEV	MNP	Clark
1	0.5238	0.0917	0.0909	0.1044	0.0071	0.0523	0.0511	0.0508	0.0685	0.0476	0.0464	0.0461	0.0675
2	0.5238	0.0950	0.0909	0.5119	0.1726	0.0703	0.0668	0.0636	0.0731	0.0639	0.0607	0.0569	0.0685
3	0.5238	0.0983	0.0909	0.9042	0.5415	0.0869	0.0857	0.0761	0.0799	0.0790	0.0779	0.0672	0.0718
4	0.5238	0.5025	0.5000	0.1044	0.0109	0.2756	0.2720	0.2715	0.2523	0.2505	0.2473	0.2464	0.2465
5	0.5238	0.5122	0.5000	0.5119	0.2620	0.3184	0.3110	0.3094	0.2896	0.2895	0.2828	0.2793	0.2713
6	0.5238	0.5215	0.5000	0.9042	0.8175	0.3486	0.3466	0.3477	0.3451	0.3169	0.3151	0.3127	0.3107
7	0.5238	0.9099	0.9091	0.1044	0.0071	0.4805	0.4794	0.4793	0.4255	0.4368	0.4359	0.4357	0.4062
8	0.5238	0.9130	0.9091	0.5119	0.1702	0.4921	0.4903	0.4913	0.4838	0.4473	0.4457	0.4462	0.4405
9	0.5238	0.9160	0.9091	0.9042	0.5282	0.4988	0.4983	0.5020	0.5006	0.4534	0.4530	0.4558	0.4546
10	0.6667	0.0991	0.0909	0.1375	0.0127	0.0714	0.0695	0.0696	0.0750	0.0357	0.0347	0.0337	0.0657
11	0.6667	0.1304	0.0909	0.5850	0.2404	0.1111	0.1065	0.1047	0.1059	0.0556	0.0533	0.0458	0.0613
12	0.6667	0.1597	0.0909	0.9260	0.6228	0.1458	0.1444	0.1381	0.1380	0.0729	0.0722	0.0561	0.0617
13	0.6667	0.5238	0.5000	0.1375	0.0192	0.3636	0.3585	0.3593	0.2832	0.1818	0.1792	0.1768	0.2265
14	0.6667	0.6000	0.5000	0.5850	0.3455	0.4444	0.4369	0.4425	0.4193	0.2222	0.2185	0.2099	0.2123
15	0.6667	0.6552	0.5000	0.9260	0.8593	0.4912	0.4896	0.5035	0.5000	0.2456	0.2448	0.2361	0.2354
16	0.6667	0.9167	0.9091	0.1375	0.0124	0.6154	0.6139	0.6144	0.5107	0.3077	0.3069	0.3063	0.3349
17	0.6667	0.9375	0.9091	0.5850	0.2140	0.6349	0.6334	0.6362	0.6289	0.3175	0.3167	0.3158	0.3152
18	0.6667	0.9500	0.9091	0.9260	0.5184	0.6437	0.6434	0.6483	0.6469	0.3218	0.3217	0.3219	0.3212

for small probabilities as high as 50 percent. This approximation generally follows the pattern of overpredicting small probabilities and underpredicting large ones. For this example the HEBA and TEV models provide better approximations to MNP than the Clark method.

Table 5.1 includes the nonsimilarity measure θ from the TEV model and $1 - \rho$ from the MNP model. Each measure lies in the unit interval, with smaller values corresponding to a greater degree of similarity. There is no simple relationship between the two scales.

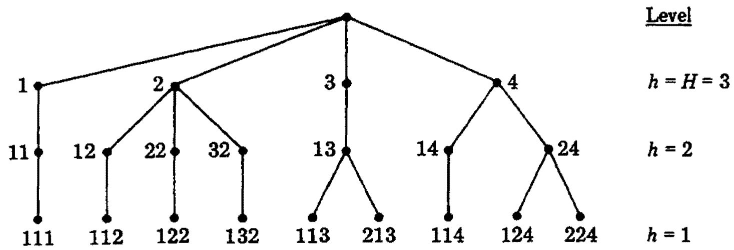
5.17 Estimation of Tree Extreme Value Models

The tree extreme value (TEV) model introduced in section 5.16 generalizes the functional form of the Luce, or multinomial logit, model to accommodate the patterns of interalternative substitution found in preference tree similarity structures, while retaining many of the computatonal advantages of the MNL model. In particular the TEV model can be written as a nested sequence of multinomial logit models, and consistent parameter estimates can be obtained from a sequence of MNL estimators.

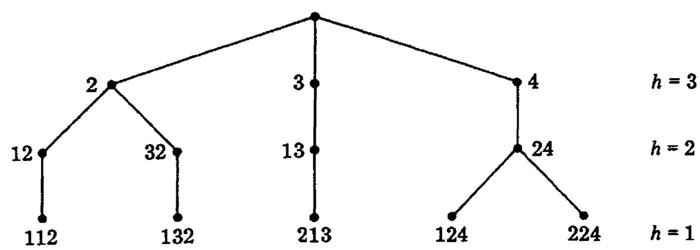
For econometric analysis it is convenient to introduce an index notation for the TEV form and write it as a nested multinomial logit (NMNL) model.

Suppose a tree has nodes at H levels, indexed $h = 1, \ldots, H$, with H denoting the stem of the tree; see figure 5.4. A node at level h in the tree is indexed by a pair (i_h, σ_h), where $\sigma_h = (i_{h+1}, \ldots, i_H)$ is the index of the adjacent node at level $h + 1$ in the tree. Thus the elemental alternatives, at level 1 in the tree, are indexed by vectors $(i_1, \ldots, i_H)$, while the alternative nodes at level H are indexed by integers i_H. For a choice set $\mathbf{B}$ let $\mathbf{B}_{\sigma_h}$ denote the set of elemental alternatives contained in $\mathbf{B}$ which are in the branch of the tree below node σ_h. We shall also use $\mathbf{B}_{\sigma_h}$ to denote the set of indices i_h such that (i_h, σ_h) is a node immediately below σ_h in the preference tree pruned to the set of elemental alternatives in $\mathbf{B}$; the interpretation will be clear from the context. Note that $\mathbf{B}_{\sigma_H} \equiv \mathbf{B}$, $\mathbf{B}_{\sigma_0}$ is the singleton σ_0 when $\sigma_0 \in \mathbf{B}$, and $\mathbf{C} = \mathbf{B}_{\sigma_h}$ can be interpreted as a choice set whose alternatives are confined to the branch below the node σ_h, with $\mathbf{C}_{\sigma_k} = \mathbf{B}_{\sigma_k}$ for $k \leq h$ and $\mathbf{C} = \mathbf{C}_{\sigma_k} = \mathbf{C}_{\sigma_H}$ for $k > h$.

Let $\mathbf{x}^h_{i_h\sigma_h}$ denote the vector of observed variables common to the alternatives below node $i_h\sigma_h$, and let γ^h be a commensurate vector of taste weights. Associate with alternative σ_0 the scale value



Preference tree



Decision tree pruned to $B = \{112, 132, 213, 124, 224\}$

$B_1 = \phi$
$B_2 = \{112, 132\}$
$B_3 = \{213\}$
$B_4 = \{124, 224\}$

$B_{11} = B_{22} = B_{14} = \phi$
$B_{12} = \{112\}$
$B_{13} = \{213\}$
$B_{24} = \{124, 224\}$
$B_{32} = \{132\}$

Figure 5.4
Index notation for preference and decision trees

$$v_{\sigma_0} = \exp\left[\sum_{h=1}^{H} \gamma^h \cdot \mathbf{x}^h_{\sigma_{h-1}}\right], \tag{5.56}$$

where $\sigma_0 \in \mathbf{B}_{\sigma_{h-1}}$. Let $\theta_{\sigma_1}, \ldots, \theta_{\sigma_H}$ be dissimilarity parameters at the nodes on the path leading to σ_0, with

$$0 < \theta_{\sigma_1} \leqq \theta_{\sigma_2} \leqq \ldots \leqq \theta_{\sigma_H} \equiv 1. \tag{5.57}$$

Then (5.49) and (5.50) imply

$$Q(\mathbf{B}_{\sigma_0} \mid \mathbf{B}_{\sigma_1}, \mathbf{s}) = \frac{\exp[\mathbf{x}^h_{i_1\sigma_1}\gamma^1/\theta_{\sigma_1}]}{\sum_{i'_1 \in \mathbf{B}_{\sigma_1}} \exp[\mathbf{x}^h_{i'_1}\gamma^1/\theta_{\sigma_1}]}. \tag{5.58}$$

For $\sigma_1 = i_2\sigma_2$ define

$$y_{i_2\sigma_2} = \ln \sum_{i_1 \in \mathbf{B}_{\sigma_1}} \exp\left[\mathbf{x}^h_{i_1\sigma_1}\frac{\gamma^1}{\theta_{\sigma_1}}\right]. \tag{5.59}$$

Recursive application of (5.49) and (5.50) for $h = 2, \ldots, H$ yields

$$Q[\mathbf{B}_{i_h\sigma_h} \mid \mathbf{B}_{\sigma_h}, \mathbf{s}] = \frac{\exp[\mathbf{x}^h_{i_h\sigma_h}\gamma^h/\theta_{\sigma_h} + y_{i_h\sigma_h}\theta_{i_h\sigma_h}/\theta_{\sigma_h}]}{\sum_{i \in \mathbf{B}_{\sigma_h}} \exp[\mathbf{x}^h_{i\sigma_h}\gamma^h/\theta_{\sigma_h} + y_{i\sigma_h}\theta_{i\sigma_h}/\theta_{\sigma_h}]} \tag{5.60}$$

and

$$y_{i_{h+1}\sigma_{h+1}} = \ln \sum_{i \in \mathbf{B}_{\sigma_h}} \exp\left[\frac{\mathbf{x}^h_{i\sigma_h}\gamma^h}{\theta_{\sigma_h}} + \frac{y_{i\sigma_h}\theta_{i\sigma_h}}{\theta_{\sigma_h}}\right]. \tag{5.61}$$

The expression $y_{i_h\sigma_h}$ is termed the inclusive value of the branch below node $i_h\sigma_h$. A necessary and sufficient condition for this model to be consistent with GEV is that (5.57) hold, or equivalently that the coefficient of each inclusive value, $\theta_{i_h\sigma_h}/\theta_{\sigma_h}$, lie in the unit interval.

For some estimation methods, it is convenient to introduce the notation

$$\boldsymbol{\beta}^h_{\sigma_h} = \left[\frac{\gamma^h}{\theta_{\sigma_h}}, \frac{\theta_{1\sigma_h}}{\theta_{\sigma_h}}, \ldots, \frac{\theta_{m\sigma_h}}{\theta_{\sigma_h}}\right], \tag{5.62}$$

where $\mathbf{B}_{\sigma_h} = \{1, \ldots, m\}$, and commensurately

$$\mathbf{z}^h_{i_h\sigma_h} = [\mathbf{x}^h_{i_h\sigma_h}, 0, \ldots, 0, y_{i_h\sigma_h}, 0, \ldots, 0]. \tag{5.63}$$

Then the model can be written

$$Q[i_h \mid \sigma_h, \mathbf{z}^h_{\cdot\sigma_h}, \boldsymbol{\beta}^h_{\sigma_h}] = \frac{\exp[\mathbf{z}^h_{i_h\sigma_h}\boldsymbol{\beta}^h_{\sigma_h}]}{\sum_{i \in \mathbf{B}_{\sigma_h}} \exp[\mathbf{z}^h_{i\sigma_h}\boldsymbol{\beta}^h_{\sigma_h}]}; \tag{5.64}$$

$$y_{\sigma_h} = \ln \sum_{i \in \mathbf{B}_{\sigma_h}} \exp[\mathbf{z}^h_{i\sigma_h} \boldsymbol{\beta}^h_{\sigma_h}].$$

The social surplus function for this PCS, from (5.36) and either (5.60) or (5.64), is

$$\begin{aligned} y_{\sigma_H} &= \ln \sum_{i_H \in \mathbf{B}} \exp[\mathbf{x}^H_{i_H} \gamma^H + y_{i_H} \theta_{i_H}] \\ &= \ln \sum_{i_H \in \mathbf{B}} \exp[\mathbf{z}^H_{i_H} \boldsymbol{\beta}^H]. \end{aligned} \tag{5.65}$$

These formulae use the conditions that σ_H is empty and $\theta_{\sigma_H} = 1$.

It should be noted that the form (5.60) implies that the coefficients of variables other than the inclusive values will differ across nodes at the same level in the tree by scale factors inversely proportional to the dissimilarity coefficients for these nodes.[37]

5.18 Sequential Estimation

Suppose a sample of T independent observations on choices from a set $\mathbf{B}$ is observed. Let $\mathbf{z}^{ht}_{\sigma_{h-1}}$ denote the vector of variables for observation t associated with h-level node σ_{h-1}. Let $m_{\sigma_{h-1}t}$ equal the number of times the choice at observation t lies in the branch below node σ_{h-1}. (Repetitions at an observation are permitted but not required.)

The conditional log likelihood of the observed transitions from the node σ_h at level $h + 1$ in the tree is

$$L_{\sigma_h} = \sum_t \sum_{i_h \in \mathbf{B}_{\sigma_h}} m_{i_h\sigma_h t} \ln Q[\mathbf{B}_{i_h\sigma_h} \mid \mathbf{B}_{\sigma_h}, \mathbf{z}^{ht}_{\cdot\sigma_h}, \boldsymbol{\beta}^h_{\sigma_h}], \tag{5.66}$$

and from all the nodes at level $h + 1$ is

$$L^h = \sum_{\sigma_h} L_{\sigma_h}. \tag{5.67}$$

The unconditional (or full information) log likelihood of the sample is

$$L = \sum_{h=1}^{H} L^h. \tag{5.68}$$

37. When all the variables in $x_{\cdot\sigma_h}$ enter as interactions with a dummy variable for node σ_h, and hence are specific to node σ_h, the parameter restrictions implied by (5.60) are satisfied trivially.

Direct maximization of the full information log likelihood yields efficient estimators of the parameters $\gamma^1, \ldots, \gamma^H$ and θ_{σ_h}, $h = 1, \ldots, H$, under standard regularity conditions.[38] While this approach presents a few technical problems, it appears to be practical when the preference tree is not too complex.[39]

An alternative estimation procedure is sequential: estimate $\boldsymbol{\beta}^1_{\sigma_1}$ by maximizing L_{σ_1}, use the estimated value of $\boldsymbol{\beta}^1_{\sigma_1}$ to compute the inclusive value variable at level 2, then estimate $\boldsymbol{\beta}^2_{\sigma_2}$ by maximizing L_{σ_2}, conditioned on the estimate of $\boldsymbol{\beta}^1_{\sigma_1}$, and so on. At each step one is estimating an MNL model by the maximum likelihood method, a standard problem for which fast and reliable algorithms exist. Since the estimation at step h involves only the parameter subvector $\boldsymbol{\beta}^h_{\sigma_h}$ rather than the full parameter vector $(\boldsymbol{\beta}^h_{\sigma_h} : h = 1, \ldots, H)$, computation costs are further reduced. Amemiya (1976) has pointed out that the use of estimators from previous steps to construct variables in the conditional likelihood at step h modifies the asymptotic covariance matrix of the estimators of $\boldsymbol{\beta}^h_{\sigma_h}$, so that standard errors produced by conventional MNL programs are incorrect. Computational formulae for the corrected covariances are given in section 5.22.

The sequential estimation procedure may be considerably less efficient than full information maximum likelihood estimation, particularly where the first-stage conditioning selects a small subsample with limited variation in some explanatory variables. A hybrid procedure is to use sequential estimation to obtain consistent estimators and then carry out one Berndt-Hausman-Hall-Hall (1974) step for the full-information log likelihood function to obtain efficient estimates. The gradients and hessians required for sequential or full-information maximum likelihood estimation are given in section 5.22.

5.19 An Application

The author and his associates have investigated work-trip choice among four travel modes (auto alone, bus, rapid transit, carpool) in the San Francisco Bay Area; see McFadden (1974), Train (1978), Train and McFadden (1978), and McFadden, Talvitie, and Associates (1977). We

38. See, for example, the conditions given by Manski and McFadden in chapter 1.

39. Unlike the simple MNL model the full information log likelihood function is not concave in parameters, complicating the numerical analysis problem of seeking and verifying a global maximum. In particular, the function is highly nonlinear in θ_{σ_h} near zero, creating problems of overflow and roundoff.

consider here estimation of nested multinomial logit models for several preference tree structures used for a sample taken in 1975 of 616 commuters. Table 5.2 gives estimates of the Luce (MNL) model. Estimates of alternative two-level trees, obtained by the sequential method, are given in table 5.3. Two of these structures have also been estimated by Cosslett (1978), using the full-information method; his results are reproduced in table 5.4.

The MNL model in table 5.2 indicates that commuters are adverse to time and cost of travel, with access (walk plus wait) time valued at 146 percent of the decision maker's wage, and on-vehicle time at 30 percent of the wage. Specification analysis of alternative models (McFadden, Talvitie, and Associates 1977) suggests that improved models are obtained by disaggregating alternatives by access mode, allowing interactions of on-vehicle times with mode dummies (to capture differential degrees of unpleasantness of time on different modes), and including socioeconomic and auto access variables in interaction with the alternatives. Assessment of the results given here on tree preference structures should be made with the limitations of the basic variable specification in mind.

Table 5.3 considers the seven possible two-level preference trees for mode choice. The names given these trees suggest aspects which when used in making similarity judgments will yield these trees; for example, "own auto access" is an attribute of the drive-alone mode and (because access is normally by auto) the rail mode, but not the remaining modes. The parameter estimates are obtained by the sequential procedure; corrected standard errors are given, using the formulae of section 5.22.[40] Variable definitions are as in table 5.2, except that "left-branch dummy" indicates a second-stage dummy variable and "inclusive value" indicates the variable defined in (5.59). For the NMNL model to be consistent with GEV or AIRUM, it is necessary that the coefficient of inclusive value lie in the unit

40. The coefficients of cost/wage, on-vehicle time, access time, and the identified mode-specific dummies are estimated in the first stage. Then inclusive values are calculated at these coefficients. In models 2, 5, and 6, there are two inclusive value coefficients, which in these models are constrained to be equal. The coefficients of inclusive value and the left-branch dummy are estimated in the second stage. Then corrected standard errors are calculated using the formulae of section 5.22. The magnitude of the correction is indicated by the following list of standard errors for the coefficient of inclusive value:

Model	2	3	4	5	6	7	8
Correct SE	0.109	0.217	0.270	0.162	0.446	0.283	0.304
Uncorrected SE	0.091	0.102	0.178	0.153	0.176	0.151	0.266

Table 5.2
An MNL model for travel mode choice (model a)[a]

Variable	Symbol	Parameter estimate (standard error)
Cost/wage, in minutes per round trip	C/W	−0.037 (0.006)
On-vehicle time, minutes per round trip	OVT	−0.010 (0.009)
Access time, minutes per round trip	AT	−0.054 (0.010)
Auto alone dummy	DA	−0.03 (0.37)
Rail transit dummy	DR	−1.06 (0.28)
Carpool dummy	DC	−1.74 (0.37)
Log likelihood[b]		−505.10

[a]The alternatives and sample for this and following models are

Mode	Number	Percent share
A: Auto, driven alone	378	61.4
B: Bus	68	11.0
R: Rail rapid transit	33	5.4
C: Carpool	137	22.2
	616	

[b]The log likelihood with dummies only is −567.60.

Table 5.3
Nested MNL models of travel mode choice

Model	2	3	4	5	6	7	8
Preference tree	*A R C B*	*B R C A*	*A B R C*	*A B R C*	*A C B R*	*A R C B*	*A B C R*
	Own auto access	Drive alone distinct	Carpool distinct	Schedule conven-ience	Transit distinct	Bus distinct	Rail distinct
Parameter estimates[a] (standard errors)							
C/W	−0.066	−0.051	−0.037	−0.017	−0.024	−0.040	−0.026
	(0.011)	(0.019)	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
OVT	−0.018	−0.021	−0.063	−0.022	−0.022	0.002	−0.014
	(0.011)	(0.011)	(0.010)	(0.011)	(0.021)	(0.015)	(0.010)
AT	−0.052	0.053	−0.053	−0.067	−0.047	−0.058	−0.056
	(0.011)	(0.011)	(0.012)	(0.014)	(0.022)	(0.023)	(0.011)
AD	1.90		0.34	−0.45	1.20	0.90	−0.39
	(0.54)		(0.45)	(0.46)	(0.31)	(0.75)	(0.41)
RD		−1.33	0.83	1.22	−1.08		
		(0.33)	(0.45)	(0.54)	(0.68)		
CD	−1.81	−1.85				−0.90	−1.83
	(0.44)	(0.44)				(0.83)	(0.40)
Left-branch dummy	−0.30	−1.12	0.92	1.43	3.01	−1.37	0.94
	(0.27)	(0.34)	(0.27)	(0.28)	(0.77)	(0.70)	(0.45)
Inclusive value	0.47	0.67	0.50	0.51	1.62	1.25	1.27
	(0.09)	(0.22)	(0.18)	(0.16)	(0.45)	(0.28)	(0.30)
Log likelihood[b]	−501.83	−502.02	−502.55	−510.32	−502.89	−503.26	−505.05

[a]For definitions of variables and modes, see table 5.2.
[b]The log likelihood with dummies only is −567.60.

interval.[41, 42] In models 2, 5, and 6, the coefficients of inclusive values in the two branches are constrained to be the same. This is not required by the TEV model; however, differing values can be accommodated in the sequential estimation procedure only by imposing a nonlinear constraint that the coefficients in one branch be a scalar multiple of the coefficients in the other branch, or else by treating the coefficients in the two branches as independent.

The alternative models 1 through 8 yield coefficients of cost, on-vehicle time, and access time of expected sign.[43] There is considerable variation between the models in the magnitude of coefficients, with models 2 and 3 implying a sharper discrimination among costs than the remainder. Estimated values of on-vehicle time range from 27 percent of the wage in model 2 to 92 percent in model 6. Extimated values of access time range from 79 percent of the wage in model 2 to 394 percent in model 5, with most values in the 100 to 200 percent range. Thus the value of time estimates are quite sensitive to model specification.

Full information maximum likelihood estimates of models 2 and 6 have been calculated by Cosslett (1978), and are given in table 5.4. In these models the coefficients of inclusive value in the two branches are allowed to differ. Thus the log likelihood is larger for these estimates both because of full model maximization and because of an additional parameter. Cosslett's estimates of model 2 do not reduce the log likelihood substantially compared to sequential estimation. However, there are substantial changes in coefficient values, implying a sensitivity to estimation method not reflected in the asymptotic standard errors. Thus caution should be exercised in interpreting these coefficients. In the FIML estimates of model 2, the coefficients of inclusive values in the two branches are not significantly different.[44] FIML estimation of model 6 results in a sub-

41. The inclusive value coefficients in models 2, 4, and 5 are significantly different from one at the 1 percent confidence level, and in model 3 at the 15 percent confidence level. In all four models the coefficient of inclusive value is significantly different from zero at the 1 percent level. It should be remembered that these tests are not independent.

42. The test statistic $T = 2[L(k) - L(1)]$, where $L(k)$ is the log likelihood for model k, has an asymptotic distribution satisfying $\mathrm{Prob}[T \leqq t] \geqq \alpha$ for $X_1^2(\alpha) = t$. Hence $T > X_1^2(\alpha)$ implies that the null hypothesis model 1 holds can be rejected with significance level at least α. By this criterion model 1 is rejected at least at the 5 percent level in (nonindependent) tests against models 2, 3, 4, and 6.

43. The coefficient of on-vehicle time in model 7 is reversed in sign but insignificant.

44. Under the null hypothesis that these coefficients are the same, the difference in the estimated coefficients is asymptotically normal with mean zero and variance equal to left coefficient variance + right coefficient variance − 2 (covariance of coefficients).

Table 5.4
Full information estimates of MNML models

Model	2	6
Preference tree	Own auto access	Transit distinct
Parameter estimates[a] **(standard errors)**		
C/W	−0.029 (0.006)	−0.056 (0.009)
OVT	−0.007 (0.006)	−0.015 (0.010)
AT	−0.032 (0.010)	−0.055 (0.013)
AD	0.37 (0.29)	0.09 (0.47)
RD	−0.30 (0.31)	1.21 (0.38)
CD	−1.11 (0.34)	3.48 (0.66)
Left inclusive value	0.48 (0.14)	2.60 (0.42)
Right inclusive value	0.59 (0.16)	1.35 (0.46)
Log likelihood[b]	−501.1	−491.1

Source: From Cosslett (1978).
[a]For definitions of variables and modes, see table 5.2.
[b]The log likelihood with dummies only is −567.60.

stantial rise in log likelihood relative to sequential estimation, again with substantial changes in parameters. For this model the coefficients of the left and right inclusive prices are significantly different, and the left coefficient is significantly greater than one. This may indicate a failure of the AIRUM specification, or may be a consequence of shortcomings in the variable specification in the model or measurement problems associated with the carpool alternative.[45]

Model 6 fitted by the FIML method yields the highest log likelihood among the models investigated. [46] This may indicate a failure of the AIRUM specification; however, more extensive FIML estimation of alternative preference trees with a more realistic variable specification would be required before a conclusion in either direction could be drawn with confidence.

The econometric models of probabilistic choice developed in this paper permit much more general patterns of similarities between alternatives than does the commonly used MNL model, while remaining reasonably practical for estimation and forecasting. The application above suggests that these models can provide significantly better fits than the MNL models. Sequential estimation of the NMNL model is practical even for relatively large and complex trees, while the FIML method is practical for problems of moderate size. The relatively large differences in coefficient estimates obtained by the two methods suggests a need for further research on the numerical and statistical properties of these methods. Finally, the numerical example in section 5.16 suggests that the MNP, HEBA, and TEV functional forms, when restricted to the same numbers of parameters, permit closely comparable fits to data generated by various patterns of similarities.

45. See McFadden, Talvitie, et al. (1977). It should be noted that, while a negative coefficient of inclusive value leads to a local failure of the GEV conditions, a coefficient of an inclusive value exceeding one will fail to satisfy GEV only for some values of the variables. Thus it is possible that an empirical fit yielding a coefficient greater than one will be consistent with GEV over the range of the data and can be combined with a second function outside the range of the data to yield a system that satisfies GEV globally. However, this chapter has not attempted to develop a test for local consistency with GEV at the observations, or for consistency with some function that satisfies GEV globally.

46. FIML estimates have not been calculated for models 3, 4, 5, 7, or 8, and no estimates have been calculated for the twelve possible three-level preference trees.

5.20 Appendix: Normalization in MNL and MNP Models

Consider the MNL model

$$P(i \mid \mathbf{z_B}, \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}\cdot\mathbf{z}_i}}{\sum_{j\in\mathbf{B}} e^{\boldsymbol{\beta}\cdot\mathbf{z}_j}} \tag{5.69}$$

and the MNP model

$$P(i \mid \mathbf{z_B}, \boldsymbol{\beta}, \mathbf{A}) = \mathrm{Prob}[\boldsymbol{\alpha}'\mathbf{z}_i \geqq \boldsymbol{\alpha}'\mathbf{z}_j \text{ for } j \in \mathbf{B}] \tag{5.70}$$

with $\boldsymbol{\alpha} \sim N(\boldsymbol{\beta}, \mathbf{AA}')$. The parameters of these models are usually fitted by iterative maximum likelihood algorithms. A practical problem in computation is that the domain of the parameters is unbounded, making it difficult to detect unbounded maxima or avoid false solutions. This problem can be avoided by a normalization which compactifies the parameter space.

Consider first the MNL model. Let $\boldsymbol{\beta}^{(k)}$ denote the vector obtained at iteration (k), and define

$$\lambda^{(k)} = [1 + \boldsymbol{\beta}^{(k)} \cdot \boldsymbol{\beta}^{(k)}]^{-1/2} > 0 \quad \text{and} \quad \bar{\boldsymbol{\beta}}^{(k)} = \lambda^{(k)}\boldsymbol{\beta}^{(k)}. \tag{5.71}$$

Then $(\lambda^{(k)}, \bar{\boldsymbol{\beta}}^{(k)})$ lies in the unit sphere and has a limit point $(\lambda^*, \bar{\boldsymbol{\beta}}^*)$. If $\lambda^* > 0$, then the likelihood attains a maximum at $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}^*/\lambda^*$. If $\lambda^* = 0$, then the likelihood has no finite maximum, and choice can be explained by nonstochastic maximization of $\bar{\boldsymbol{\beta}}^* \cdot \mathbf{z}_i$, with $\bar{\boldsymbol{\beta}}^* \cdot \bar{\boldsymbol{\beta}}^* = 1$. Termination of the iterative algorithm for sufficiently small changes in $(\lambda^{(k)}, \bar{\boldsymbol{\beta}}^{(k)})$ yields a reliable convergence criterion.

Consider the MNP model, and let $(\boldsymbol{\beta}^{(k)}, \mathbf{A}^{(k)})$ denote the parameter values at iteration k. Suppose that as a result of normalization $\boldsymbol{\beta}^{(k)} \cdot \boldsymbol{\beta}^{(k)} + \mathrm{tr}[\mathbf{A}^{(k)}\mathbf{A}^{(k)\prime}] = 1$, and suppose that the diagonal elements of $\mathbf{A}^{(k)}$ are positive. Fix the element $A_{11}^{(k)}$ (and other normalizations as necessary for nonsingularity), and iterate to new values $\tilde{\boldsymbol{\beta}}^{(k+1)}$ and $\tilde{\mathbf{A}}^{(k+1)}$, constraining the algorithm to keep the diagonal of $\tilde{\mathbf{A}}^{(k+1)}$ positive. Define

$$(\boldsymbol{\beta}^{(k+1)}, \mathbf{A}^{(k+1)})$$
$$= (\tilde{\boldsymbol{\beta}}^{(k+1)}, \tilde{\mathbf{A}}^{(k+1)}) \cdot [\tilde{\boldsymbol{\beta}}^{(k+1)} \cdot \tilde{\boldsymbol{\beta}}^{(k+1)} + \mathrm{tr}\, \tilde{\mathbf{A}}^{(k+1)} \tilde{\mathbf{A}}^{(k+1)\prime}]^{-1/2}. \tag{5.72}$$

Then the parameter values at iteration $k + 1$ again lie in the unit sphere and have the diagonal of $\mathbf{A}^{(k+1)}$ positive. The sequence $(\boldsymbol{\beta}^{(k)}, \mathbf{A}^{(k)})$ has a (possibly

nonunique) limit point $(\boldsymbol{\beta}^*, \mathbf{A}^*)$ at which the likelihood function has a local maximum. Note that $\mathbf{A}^*$ may be degenerate, in which case the likelihood is defined on the linear subspace spanned by $\mathbf{A}^*$. This poses no difficulty in the algorithm or in the interpretation of $(\boldsymbol{\beta}^*, \mathbf{A}^*)$. Termination of the iterative algorithm for sufficiently small changes in $(\boldsymbol{\beta}^{(k)}, \mathbf{A}^{(k)})$ should identify (local) maxima.

5.21 Appendix: Computational Formulas for a Simple Model

Consider the preference tree of figure 5.3, with choice probabilities satisfying the trinary condition (5.53), and the HEBA, TEV, or MNP models. The HEBA model specification (5.54) can be inverted to

$$v_1 = \frac{r_{13} - r_{23}}{1 - r_{21}}, \quad v_2 = \frac{r_{13} - r_{23}}{r_{12} - 1},$$

$$v_3 = 1, \qquad v_{12} = \frac{r_{23} - r_{13} r_{21}}{1 - r_{21}}, \tag{5.73}$$

where $r_{ij} = p_{ij}/p_{ji}$. The multinomial probabilities are then determined from the binary probabilities using the last equations in (5.54).

The TEV model with the generating function (5.55) has

$$p_{12} = \frac{y_1^{1/\theta}}{y_1^{1/\theta} + y_2^{1/\theta}},$$

$$p_{13} = \frac{y_1}{y_2 + y_3},$$

$$p_{23} = \frac{y_2}{y_2 + y_3}. \tag{5.74}$$

Inverting

$$y_1 = r_{13}, \quad y_2 = r_{23}, \quad y_3 = 1,$$

$$\theta = \frac{\ln y_1/y_2}{\ln r_{12}}. \tag{5.75}$$

The trinary condition requires that $r_{12} > 1$ $(y_1 > y_2)$ imply $r_{12} > r_{13}/r_{23} > 1$, or $(y_1/y_2)^{1/\theta} > y_1/y_2 > 1$. This holds if and only if $0 < \theta < 1$. From (5.55) the multinomial choice probabilities satisfy

$$
\begin{aligned}
P_3 &= [1 + (y_1^{1/\theta} + y_2^{1/\theta})^\theta]^{-1}, \\
P_1 &= p_{12}(1 - P_3), \\
P_2 &= p_{21}(1 - P_3).
\end{aligned} \tag{5.76}
$$

Consider the MNP model with a random utility vector $(u_1, u_2, u_3) \sim N((\mu_1, \mu_2, 0), \Omega)$, where

$$
\Omega = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 \\ \sigma_{21} & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}.
$$

The condition $\mu_3 = 0$ is a normalization; the independence of u_3 and (u_1, u_2), and $\sigma_{12} > 0$, is assumed in correspondence with the tree structure in figure 5.3. The binary choice probabilities then satisfy

$$
\begin{aligned}
p_{13} &= \Phi\left[\frac{\mu_1}{\sqrt{\sigma_{11} + \sigma_{33}}}\right], \\
p_{23} &= \Phi\left[\frac{\mu_2}{\sqrt{\sigma_{22} + \sigma_{33}}}\right], \\
p_{12} &= \Phi\left[\frac{\mu_1 - \mu_2}{\sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}}\right].
\end{aligned} \tag{5.77}
$$

The trinary condition requires that $r_{12} > 1$ $(\mu_1 > \mu_2)$ imply

$$
\frac{r_{13}}{r_{23}} > 1 \Rightarrow \left(\frac{\mu_1}{\sqrt{\sigma_{11} + \sigma_{33}}} > \frac{\mu_2}{\sqrt{\sigma_{22} + \sigma_{33}}}\right),
$$

and hence $\sigma_{11} = \sigma_{22}$. Then by standardizing the variance of utility differences, one can show that there is no loss of generality in imposing as normalizing restrictions $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1/2$ and $\sigma_{12} = \rho/2$, with $0 < \rho < 1$. Hence

$$
p_{13} = \Phi(\mu_1), \quad p_{23} = \Phi(\mu_2), \quad p_{12} = \Phi\left[\frac{\mu_1 - \mu_2}{\sqrt{1 - \rho}}\right]. \tag{5.78}
$$

After some manipulation the multinomial choice probabilities can be written

$$P_1 = \int_{(\mu_2-\mu_1)/\sqrt{1-\rho}}^{\infty} \phi(t)\Phi\left[\frac{2\mu_1}{\sqrt{3+\rho}} + t\sqrt{\frac{1-\rho}{3+\rho}}\right]dt, \tag{5.79}$$

$$P_2 = \int_{(\mu_1-\mu_2)/\sqrt{1-\rho}}^{\infty} \phi(t)\Phi\left[\frac{2\mu_2}{\sqrt{3+\rho}} + t\sqrt{\frac{1-\rho}{3+\rho}}\right]dt.$$

The Clark approximation to these probabilities is

$$P_1 \approx \Phi\left[\frac{\mu_1 - p_{13}\mu_2 - \phi(\mu_2)}{[1 - \rho p_{13} + \mu_2^2 p_{13}p_{31} + \mu_2\phi(\mu_1) - \phi(\mu_1)^2 - 2\mu_2 p_{13}\phi(\mu_2)]^{1/2}}\right],$$

$$P_2 \approx \Phi\left[\frac{\mu_2 - p_{23}\mu_1 - \phi(\mu_1)}{[1 - \rho p_{23} + \mu_1^2 p_{23}p_{32} + \mu_1\phi(\mu_2) - \phi(\mu_2)^2 - 2\mu_1 p_{23}\phi(\mu_1)]^{1/2}}\right]. \tag{5.80}$$

5.22 Appendix: Computational Formulas for the Nested Multinomial Logit Model

This section first describes a method for calculating asymptotic covariance matrices for sequential or full-information maximum likelihood estimates of an NMNL model. Second, formulae are given for the required derivatives. Finally, there is some discussion of algorithms for implementing the computation.

The log likelihood function for NMNL can be written in the general schematic form

$$\begin{aligned} L^0(\mathbf{x}_1, \ldots, \mathbf{x}_m \mid \psi_1, \ldots, \psi_m) \equiv {} & L^1(\mathbf{x}_1 \mid \mathbf{x}_2, \ldots, \mathbf{x}_m, \psi_1) \\ & + L^2(\mathbf{x}_2 \mid \mathbf{x}_3, \ldots, \mathbf{x}_m, \psi_1, \psi_2) \\ & + \ldots + L^m(\mathbf{x}_m \mid \psi_1, \ldots, \psi_m), \end{aligned} \tag{5.81}$$

where each ψ_i is a parameter vector, $\mathbf{x}_i$ is a data vector, and L^i is the conditional log likelihood associated with transitions at one level of the decision tree.

Full information maximum likelihood (FIML) estimators satisfy

$$L^0_\psi(\mathbf{x}_1, \ldots, \mathbf{x}_m \mid \hat{\psi}_1, \ldots, \hat{\psi}_m) = 0, \tag{5.82}$$

where $\psi = (\psi_1, \ldots, \psi_m)$ and $L^0_\psi = \partial L^0/\partial\psi$. A Taylor's expansion of (5.82) about the true parameter vector ψ^* yields

$$\left[\frac{1}{N} L^0_{\psi\psi}\right]\sqrt{N}\,(\hat{\psi} - \psi^*) = -\frac{1}{\sqrt{N}} L^0_\psi, \tag{5.83}$$

where N is the sample size, L^0_ψ is evaluated at ψ^*, and the rows of $L^0_{\psi\psi}$ are evaluated at points between $\hat{\psi}$ and ψ^*. Using the conventional asymptotic development for maximum likelihood estimation, one has under standard regularity conditions that $\hat{\psi}$ is consistent for ψ^*, by the law of large numbers that plim $(1/N)L^0_{\psi\psi} = \lim E(1/N)L^0_{\psi\psi} \equiv \mathbf{B}_0$, and by a central limit theorem that $(1/\sqrt{N})L^0_\psi$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\lim E(1/N)L^0_\psi L^{0\prime}_\psi \equiv \mathbf{A}_0$. Further, differentiation of the identity $\mathrm{E}e^{L^0} \equiv 1$ yields $\mathbf{A}_0 = -\mathbf{B}_0$. Then $\sqrt{N}(\hat{\psi} - \psi^*)$ is asymptotically normal with zero mean and covariance matrix $\mathbf{\Omega}_0 \equiv \mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{B}_0^{-1} \equiv \mathbf{B}_0^{-1}$.

The sequential estimation procedure first determines $\tilde{\psi}_1$, satisfying

$$L^1_{\psi_1}(\mathbf{x}_1 \mid \mathbf{x}_2, \ldots, \mathbf{x}_m, \tilde{\psi}_1) = 0, \tag{5.84}$$

and then recursively determines $\tilde{\psi}_i$, satisfying

$$L^i_{\psi_i}(\mathbf{x}_i \mid \mathbf{x}_{i+1}, \ldots, \mathbf{x}_m, \tilde{\psi}_1, \ldots, \tilde{\psi}_{i-1}, \tilde{\psi}_i) = \mathbf{0}, \tag{5.85}$$

given the previously estimated values of $\tilde{\psi}_1, \ldots, \tilde{\psi}_{i-1}$, for $i = 2, \ldots, m$. The development of the asymptotic properties of the sequential estimator parallels that of the FIML estimator. A series of Taylor's expansions of (5.84) and (5.85) yield

$$\frac{1}{N}\begin{bmatrix} L^1_{\psi_1\psi_1} & \mathbf{0} & \cdots & \mathbf{0} \\ L^2_{\psi_1\psi_2} & L^2_{\psi_2\psi_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ L^m_{\psi_1\psi_m} & L^m_{\psi_2\psi_m} & \cdots & L^m_{\psi_m\psi_m} \end{bmatrix} \begin{bmatrix} \sqrt{N}\,(\tilde{\psi}_1 - \psi^*_1) \\ \sqrt{N}\,(\tilde{\psi}_2 - \psi^*_2) \\ \vdots \\ \sqrt{N}\,(\tilde{\psi}_m - \psi^*_m) \end{bmatrix} = -\frac{1}{\sqrt{N}} \begin{bmatrix} L^1_{\psi_1} \\ L^2_{\psi_2} \\ \vdots \\ L^m_{\psi_m} \end{bmatrix}, \tag{5.86}$$

or in matrix notation

$$\mathbf{B}_N\sqrt{N}\,(\tilde{\psi} - \psi^*) = -\frac{1}{\sqrt{N}}\lambda_N. \tag{5.87}$$

Analogously to the FIML case $\mathbf{B}_N$ converges in probability to a lower block triangular matrix $\mathbf{B}_* = \lim E\mathbf{B}_N$ and, $(1/\sqrt{N})\boldsymbol{\lambda}_N$ converges in distribution to an asymptotically normal random variable with zero mean and covariance matrix $\mathbf{A}_* = \lim E(1/N)\boldsymbol{\lambda}_N\boldsymbol{\lambda}'_N$. Then $\sqrt{N}(\tilde{\psi} - \psi^*)$ is asymptotically normal with zero mean and covariance matrix $\mathbf{B}_*^{-1}\mathbf{A}_*\mathbf{B}_*^{-1}$. Computation of this covariance matrix is facilitated by noting first that at ψ^*

$$\lim E\frac{1}{N}L^i_{\psi_{i-k}\psi_i} = -\lim E\frac{1}{N}L^i_{\psi_{i-k}}L^i_{\psi_k}, \tag{5.88}$$

and second that the block triangular matrix $\mathbf{B}_*^{-1}$ can be calculated using a recursion formula which requires inverting only matrices of the order of the blocks $L^i_{\psi_i\psi_i}$.

The following paragraphs give formulae for the derivatives of the log likelihood function of the NMNL model. From (5.64) define $i = i_h$, $\sigma = \sigma_h$, and

$$\begin{aligned}\ln q(i \mid \sigma) &\equiv \ln Q[i \mid \sigma_h, \mathbf{z}^h_{\cdot\sigma_h}, \boldsymbol{\beta}_{\sigma_h}] \\ &= \mathbf{x}^h_{i\sigma}\frac{\gamma^h}{\theta_\sigma} + \frac{\theta_{i\sigma}}{\theta_\sigma}y_{i\sigma} \\ &\quad - \ln\sum_{j\in\mathbf{B}_\sigma}\exp\left(\frac{\mathbf{x}^h_{j\sigma}\gamma^h}{\theta_\sigma} + \frac{\theta_{j\sigma}}{\theta_\sigma}y_{j\sigma}\right).\end{aligned} \tag{5.89}$$

Then

$$\frac{\partial \ln q(i \mid \sigma)}{\partial \gamma^{h-r}} = \frac{(\mathbf{x}^{h-r}_{i\sigma} - \mathbf{x}^{h-r}_{\sigma})}{\theta_\sigma} \tag{5.90}$$

for $r = 0, \ldots, h-1$, where

$$\mathbf{x}^{h-r}_{\sigma} = \sum_{i_{h-r}} \cdots \sum_{i_h} q(i_{h-r}\ldots i_h \mid \sigma)\mathbf{x}^{h-r}_{i_{h-r}\cdots i_h\sigma} \tag{5.91}$$

and

$$q(i_{h-r}\ldots i_h \mid \sigma) = q(i_{h-r} \mid i_{h-r+1}\ldots i_h\sigma)\ldots q(i_h \mid \sigma). \tag{5.92}$$

Further

$$\begin{aligned}\frac{\partial \ln q(i \mid \sigma)}{\partial(1/\theta_\sigma)} &= (\mathbf{x}^h_{i\sigma} - \mathbf{x}^h_\sigma)\gamma^h + y_{i\sigma}\theta_{i\sigma} \\ &\quad - \sum_j q(j \mid \sigma)y_{j\sigma}\theta_{j\sigma},\end{aligned} \tag{5.93}$$

and

$$\frac{\partial \ln q(i \mid \sigma)}{\partial(1/\theta_{i_{h-r}\cdots i_h\sigma})} = [\delta_{i_h i} - q(i_h \mid \sigma)]\,\alpha[i_{h-r}\ldots i_h\sigma], \tag{5.94}$$

with $\delta_{ij} = 1$ if $i = j$ and

$$\begin{aligned}\alpha(i_{h-r}\ldots i_h\sigma) = q(i_{h-r}\ldots i_{h-1} \mid i_h\sigma)&\\ \cdot\frac{\theta_{i_{h-r}\cdots i_h\sigma}}{\theta_\sigma}\Big\{&\mathbf{x}^{h-r-1}_{i_{h-r}\cdots i_h\sigma}\gamma^{h-r-1}\\ &+\sum_j q(j \mid i_{h-r}\ldots i_h\sigma)y_{ji_{h-r}\cdots i_h\sigma}\theta_{ji_{h-r}\cdots i_h\sigma}\\ &- y_{i_{h-r}\cdots i_h\sigma}\theta_{i_{h-r}\cdots i_h\sigma}\Big\}.\end{aligned} \tag{5.95}$$

Consider the conditional log likelihood function for transition from node σ, which from (5.66) satisfies

$$L_\sigma = \sum_t \sum_{i \in \mathbf{B}_\sigma} m_{i\sigma t} \ln q(i \mid \sigma, t), \tag{5.96}$$

with $m_{i\sigma t}$ the number of choices of $i\sigma$ at case t and $q(i \mid \sigma, t)$ the transition probability for case t.

Then

$$\frac{\partial L_\sigma}{\partial \gamma^{h-r}} = \frac{\sum_t \left[\sum_{i \in \mathbf{B}_\sigma} m_{i\sigma t}\mathbf{x}^{h-r,t}_{i\sigma} - m_{\sigma t}\mathbf{x}^{h-r,t}_{\sigma}\right]}{\theta_\sigma}; \tag{5.97}$$

$$\frac{\partial L_\sigma}{\partial(1/\theta_\sigma)} = \sum_t \sum_{i \in \mathbf{B}_\sigma} (m_{i\sigma t} - m_{\sigma t} q(i \mid \sigma, t))[\mathbf{x}^{ht}_{i\sigma}\gamma^h + y^t_{i\sigma}\theta_{i\sigma}]; \tag{5.98}$$

$$\frac{\partial L_\sigma}{\partial[1/\theta_{i_{h-r}\cdots i_h\sigma}]} = \sum_t [m_{i_h\sigma t} - m_{\sigma t} q(i_h \mid \sigma t)]\,\alpha(i_{h-r}\cdots i_h\sigma, t). \tag{5.99}$$

The distribution of $m_{i\sigma t}$, conditioned on $m_{\sigma t}$, is multinomial, statistically independent of transitions at nodes other than σ and by assumption independent across t. The gradients (5.97) through (5.99) have zero means, when evaluated at the true parameter vector, and covariances

$$E\left[\frac{\partial L_\sigma}{\partial \gamma^{h-r}}\right]\left[\frac{\partial L_\sigma}{\partial \gamma^{h-s}}\right]' = \sum_t \sum_{i \in \mathbf{B}_\sigma} q(i \mid \sigma, t)(\mathbf{x}_{i\sigma}^{h-r,t} - \mathbf{x}_\sigma^{h-r,t}) \cdot \frac{(\mathbf{x}_{i\sigma}^{h-s,t} - \mathbf{x}_\sigma^{h-s,t})'}{\theta_\sigma^2}; \tag{5.100}$$

$$E\left[\frac{\partial L_\sigma}{\partial [1/\theta_\sigma]}\right]^2 = \sum_t m_{\sigma t} \left\{ \sum_{i \in \mathbf{B}_\sigma} q(i \mid \sigma, t)(\mathbf{x}_{i\sigma}^{ht} \gamma^h + y_{i\sigma}^t \theta_{i\sigma})^2 - \left[\sum_{i \in \mathbf{B}_\sigma} q(i \mid \sigma, t)(\mathbf{x}_{i\sigma}^{ht} \gamma^h + y_{i\sigma}^t \theta_{i\sigma}) \right]^2 \right\}; \tag{5.101}$$

$$E \frac{\partial L_\sigma}{\partial [1/\theta_{i_{h-r} \cdots i_h \sigma}]} \frac{\partial L_\sigma}{\partial [1/\theta_{i'_{h-s} \cdots i'_h \sigma}]} = \sum_t m_{\sigma t} q(i_h \mid \sigma, t)[\delta_{i_h i'_h} - q(i'_h \mid \sigma, t)] \cdot \alpha(i_{h-r} \cdots i_h \sigma, t) \alpha(i'_{h-s} \cdots i'_h \sigma, t); \tag{5.102}$$

$$E \frac{\partial L_\sigma}{\partial \gamma^{h-r}} \frac{\partial L_\sigma}{\partial [1/\theta_\sigma]} = \sum_t m_{\sigma t} \left\{ \left[\sum_i q(i \mid \sigma, t) \mathbf{x}_{i\sigma}^{h-r,t} \mathbf{x}_{i\sigma}^{h,t'} - \mathbf{x}_\sigma^{h-r,t} \mathbf{x}_\sigma^{h'} \right] \gamma^h + \sum_i q(i \mid \sigma, t) y_{i\sigma}^t \theta_{i\sigma} (\mathbf{x}_{i\sigma}^{h-r,t} - \mathbf{x}_\sigma^{h-r,t})/\theta_\sigma \right\}; \tag{5.103}$$

$$E \frac{\partial L_\sigma}{\partial \gamma^{h-r}} \frac{\partial L_\sigma}{\partial [1/\theta_{i_{h-s} \cdots i_h \sigma}]} = \sum_t m_{\sigma t} q(i_h \mid \sigma, t) \alpha(i_{h-s} \cdots i_h \sigma, t) \cdot \frac{[\mathbf{x}_{i_h \sigma}^{h-r,t} - \mathbf{x}_\sigma^{h-r,t}]}{\theta_\sigma}; \tag{5.104}$$

$$E \frac{\partial L_\sigma}{\partial [1/\theta_\sigma]} \frac{\partial L_\sigma}{\partial [1/\theta_{i_{h-s} \cdots i_h \sigma}]} = \sum_t m_{\sigma t} q(i_h \mid \sigma, t) \alpha(i_{h-s} \cdots i_h \sigma, t) \cdot \left\{ [\mathbf{x}_{i_h \sigma}^{ht} - \mathbf{x}_\sigma^{ht}] \gamma^h + y_{i_h \sigma}^t \theta_{i_h \sigma} - \sum_i q(i \mid \sigma, t) y_{i\sigma}^t \theta_{i\sigma} \right\}. \tag{5.105}$$

The covariance matrices for FIML or sequential maximum likelihood estimators are constructed from (5.100) through (5.105), using the asymptotic methods given at the beginning of this section and the definitions (5.66) through (5.68) of the conditional and unconditional log likelihood functions. The FIML log likelihood is the sum over the nodes σ in the decision tree of the terms L_σ. The gradients of L_σ and $L_{\sigma'}$ have zero covariance for $\sigma \neq \sigma'$; hence the information matrix for FIML has elements equal to the sum over σ of terms like (5.100) or (5.105). This matrix is readily estimated by substituting the FIML estimates of the parameters in (5.100) through (5.105), and its inverse yields an estimate of the asymptotic covariance matrix of the maximum likelihood estimator.

The structure of sequential estimation is simplified substantially if either the dissimilarity parameters θ_{σ_h} at level h in the tree are constrained a priori to be equal or the proportionality constraints on parameters across nodes at the same level of the tree are ignored. In the first case estimation can be carried out sequentially over the levels of the tree. At each level the parameters are estimated by applying a multinomial logit maximum likelihood procedure to the choice data for all the transitions at this level. For this procedure it is convenient to use the parameterization (5.62) with a common dissimilarity parameter θ_h at level h, writing

$$\boldsymbol{\beta}^h \equiv (\boldsymbol{\beta}_1^h, \boldsymbol{\beta}_2^h) = \left(\frac{\gamma^h}{\theta_h}, \frac{\theta_{h-1}}{\theta_h}\right) \tag{5.106}$$

$$\mathbf{z}_{i_h\sigma_h}^h = [\mathbf{x}_{i_h\sigma_h}^h, y_{i_h\sigma_h}]$$

$$Q[i_h \mid \sigma_h, \mathbf{z}_{\cdot\sigma_h}^h, \boldsymbol{\beta}^h] = \frac{\exp[\mathbf{z}_{i_h\sigma_h}^h \boldsymbol{\beta}^h]}{\sum_{i \in \mathbf{B}_{\sigma_h}} \exp[\mathbf{z}_{i\sigma_h}^h \boldsymbol{\beta}^h]}.$$

Then the log likelihood function satisfies

$$L^h = \sum_t \sum_{\sigma_h} \sum_{i \in \mathbf{B}_{\sigma_h}} m_{i\sigma_h t} \ln q(i \mid \sigma_h, t); \tag{5.107}$$

$$\frac{\partial L^h}{\partial \boldsymbol{\beta}^{h-r}} = \sum_t \sum_{\sigma_h} \lambda_r^h \left[\sum_{i \in \mathbf{B}_{\sigma_h}} \mathbf{z}_{i\sigma_h}^{h-r,t} m_{i\sigma_h t} - \mathbf{z}_{\sigma_h}^{h-r,t} m_{\sigma_h t} \right], \tag{5.108}$$

with $\lambda_0^h = 1$ and $\lambda_r^h = \beta_2^{h-r+1} \cdots \beta_2^h$ for $r > 0$; and

$$E\left[\frac{\partial L^h}{\partial \boldsymbol{\beta}^{h-r}}\right]\left[\frac{\partial L^h}{\partial \boldsymbol{\beta}^{h-s}}\right]'$$

$$= \sum_t \sum_{\sigma_h} \lambda_r^h \lambda_s^h m_{\sigma_h t}\left[\sum_{i\in \mathbf{B}_{\sigma_h}} q(i\mid \sigma_h, t)\mathbf{z}_{i\sigma_h}^{h-r,t\prime}\mathbf{z}_{i\sigma_h}^{h-s,t'} - \mathbf{z}_{\sigma_h}^{h-r,t\prime}\mathbf{z}_{\sigma_h}^{h-s,t'}\right]. \quad (5.109)$$

The parameters γ^h and θ_h can be obtained from the $\boldsymbol{\beta}$'s using the transformations

$$\theta_h = \beta_2^{h+1} \cdots \beta_2^H \quad (5.110)$$

$$\gamma^h = \boldsymbol{\beta}_1^h \theta_h.$$

Consider the case where the dissimilarity parameters are not constrained to be equal across nodes at the same level of the tree and the proportionality restrictions across these nodes on the coefficients of variables other than inclusive values are ignored. Note that failure to impose these conditions entails some loss of information, except in the case where all variables are defined by interactions with alternative or branch-specific dummies. In this case the estimation procedure using (5.107) through (5.109) can be applied to each branch separately; the formulae are modified solely by dropping the summation over σ_h.

The general case of sequential estimation, where dissimilarity parameters at each level of the tree are not constrained to be equal, and parameter restrictions across nodes are imposed, can be treated in the same framework as the FIML analysis. However, the estimation problem at each level no longer has a simple multinomial logit structure, and there appears to be little reason, except possibly the scale of the problem, to use a sequential rather than FIML estimation.

Consider algorithms for implementing the computation of sequential or FIML estimators of the NMNL model, and estimators of the covariance matrix of the estimates. Letting $L(\psi)$ denote the log likelihood function for FIML or for one step of sequential estimation, a practical and relatively efficient search algorithm chooses the direction of search from a trial parameter vector $\psi^{(k)}$, satisfying

$$\psi^{(k+1)} - \psi^{(k)} = \lambda\left[E\left[\frac{\partial L}{\partial \psi}\right]\left[\frac{\partial L}{\partial \psi}\right]'\right]^{-1}\frac{\partial L}{\partial \psi}, \quad (5.111)$$

with the right-hand side derivatives evaluated at $\psi^{(k)}$ and λ a positive step size. In a neighborhood of the optimum, this algorithm has quadratic convergence for $\lambda = 1$. The search direction will always be a line of ascent, even if the function is not locally concave. These properties are discussed further in Berndt-Hausman-Hall-Hall (1974). It is often efficient to carry out an interpolation along the direction of search, choosing λ, so that an approximate optimum is obtained and the following search direction will be nearly orthogonal. Using the calculated levels and gradients of the function at points along the direction of search which straddle the maximum permits a fairly accurate interpolation.

In the sequential estimation procedure the asymptotic covariance structure can be utilized to give recursive formulae for the covariance matrices at each stage. In the case where dissimilarity coefficients within each level of the tree are constrained to be equal, let

$$\mathbf{M}^h_{h-r,h-s} = E\left[\frac{\partial L^h}{\partial \boldsymbol{\beta}^{h-r}}\right]\left[\frac{\partial L^h}{\partial \boldsymbol{\beta}^{h-s}}\right]', \quad r, s = 0, \ldots, h-1. \tag{5.112}$$

$$\mathbf{B}_N = \frac{1}{N}\begin{bmatrix} \mathbf{M}^1_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{M}^2_{21} & \mathbf{M}^2_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \\ \mathbf{M}^H_{H1} & \mathbf{M}^H_{H2} & \cdots & \mathbf{M}^H_{HH} \end{bmatrix}, \tag{5.113}$$

and

$$\mathbf{A}_N = \frac{1}{N}\begin{bmatrix} \mathbf{M}^1_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^2_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}^H_{HH} \end{bmatrix}$$

Then the asymptotic covariance matrix of

$$\sqrt{N}\,(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \equiv \sqrt{N}\,(\tilde{\boldsymbol{\beta}}^1 - \boldsymbol{\beta}^{1*}, \ldots, \tilde{\boldsymbol{\beta}}^H - \boldsymbol{\beta}^{H*}) \tag{5.114}$$

is $\mathbf{B}_N^{-1}\mathbf{A}_N\mathbf{B}_N^{-1}$. Let

$$\mathbf{B}_N^{-1} \equiv \mathbf{C}_N = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \\ \mathbf{C}_{H1} & \mathbf{C}_{H2} & \cdots & \mathbf{C}_{HH} \end{bmatrix}. \tag{5.115}$$

Then recursion formulae for the asymptotic covariance matrix are

$$\mathbf{C}_{hh} = (\mathbf{M}^h_{hh})^{-1}; \tag{5.116}$$

$$\mathbf{C}_{h,h-r} = -\mathbf{C}_{hh} \sum_{l=h-r}^{h-1} \mathbf{M}^{h-1}_{h-1,l}\mathbf{C}_{l,h-r} \quad \text{for} \quad r > 0; \tag{5.117}$$

$$\mathbf{V}_{h,h-r} = \sum_{l=1}^{h-r} \mathbf{C}_{hl}\mathbf{M}^l_{ll}\mathbf{C}'_{h-r,l}, \tag{5.118}$$

where

$$\mathbf{B}_N^{-1}\mathbf{A}_N\mathbf{B}_N^{-1} = \begin{bmatrix} \mathbf{V}_{11} & \cdots & \mathbf{V}_{1H} \\ \vdots & & \vdots \\ \mathbf{V}_{H1} & \cdots & \mathbf{V}_{HH} \end{bmatrix}. \tag{5.119}$$

5.23 Appendix: Proof of Theorem 5.1

i. Supose an AIRUM form is given, with individual indirect utility having the form $u(i) = (y - q_i + \varepsilon_i)/\beta(\mathbf{r})$, and $\boldsymbol{\varepsilon}_{\mathbf{B}}$ distributed in the population with a cumulative distribution function $F(\boldsymbol{\varepsilon}_{\mathbf{B}})$ and density $f(\boldsymbol{\varepsilon}_{\mathbf{B}})$. Define

$$G(\mathbf{q}_{\mathbf{B}}) = \int_{-\infty}^{+\infty} [F(t + \mathbf{0}_{\mathbf{B}}) - F(t + \mathbf{q}_{\mathbf{B}})]\, dt. \tag{5.120}$$

First, we will show that G exists and is differentiable. Let F^i denote the marginal cumulative distribution function of ε_i. If $\lambda = \max_{i \in \mathbf{B}} |q_i - q'_i|$, then

$$\begin{aligned} F(t + \mathbf{q}_{\mathbf{B}}) - F(t + \mathbf{q}_{\mathbf{B}} + \lambda) &\leqq F(t + \mathbf{q}_{\mathbf{B}}) - F(t + \mathbf{q}'_{\mathbf{B}}) \\ &\leqq F(t + \mathbf{q}_{\mathbf{B}}) - F(t + \mathbf{q}_{\mathbf{B}} - \lambda), \end{aligned} \tag{5.121}$$

implying $|F(t + \mathbf{q}_{\mathbf{B}}) - F(t + \mathbf{q}'_{\mathbf{B}})| \leqq F(t + \mathbf{q}_{\mathbf{B}} + \lambda) - F(t + \mathbf{q}_{\mathbf{B}} - \lambda)$. Since F is a cumulative distribution function,

$$\begin{aligned} &F(t + \mathbf{q}_{\mathbf{B}} + \lambda) - F(t + \mathbf{q}_{\mathbf{B}} - \lambda) \\ &\qquad \leqq \sum_{i=1}^{m} [F^i(t + q_i + \lambda) - F^i(t + q_i - \lambda)]. \end{aligned} \tag{5.122}$$

For any scalar $M \geqq 0$ and positive integer K,

$$\int_{M}^{M+K\lambda} [F^i(t+q_i+\lambda) - F^i(t+q_i-\lambda)]\,dt$$

$$= \sum_{k=1}^{K} \int_{M+(k-1)\lambda}^{M+k\lambda} [F^i(t+q_i+\lambda) - F^i(t+q_i-\lambda)]\,dt$$

$$= \int_{M+(K-1)\lambda}^{M+(K+1)\lambda} F^i(t+q_i)\,dt - \int_{M-\lambda}^{M+\lambda} F^i(t+q_i)\,dt$$

$$\leqq 2\lambda\{F^i(M+(K+1)\lambda+q_i) - F^i(M-\lambda+q_i)\}. \qquad (5.123)$$

Letting $K \to +\infty$, (5.121) through (5.123) imply

$$\int_{M}^{\infty} |F(t+\mathbf{q_B}) - F(t+\mathbf{q'_B})|\,dt \leqq 2\lambda \sum_{i=1}^{m} (1 - F^i(M-\lambda+q_i)). \qquad (5.124)$$

A similar argument yields

$$\int_{-\infty}^{-M} |F(t+\mathbf{q_B}) - F(t+\mathbf{q'_B})|\,dt \leqq 2\lambda \sum_{i=1}^{m} F^i(-M+\lambda+q_i). \qquad (5.125)$$

Taking $\mathbf{q'_B} = \mathbf{0_B}$ and $M = 0$ implies

$$\int_{-\infty}^{+\infty} |F(t+\mathbf{q_B}) - F(t)|\,dt \leqq 4m \max |q_i|. \qquad (5.126)$$

Hence G defined by (5.120) exists.

For $\theta > 0$,

$$G(\mathbf{q_B}) - G(\mathbf{q_B}+\theta) = \lim_{K\to\infty} \sum_{i=-K}^{K-2} \int_{i\theta}^{(i+1)\theta} [F(t+\mathbf{q_B}+\theta) - F(t+\mathbf{q_B})]\,dt$$

$$= \lim_{K \to \infty} \left\{ \int_{(K-1)\theta}^{K\theta} F(t + q_{\mathbf{B}})\, dt - \int_{-K\theta}^{(1-K)\theta} F(t + q_{\mathbf{B}})\, dt \right\}$$

$$= \theta. \tag{5.127}$$

An analogous argument establishes (5.127) for $\theta < 0$. Hence SS 5.3 holds.

Next the differentiability of G is established. For $\mathbf{q_B} = \mathbf{q}'_{\mathbf{B}} + \theta \mathbf{q}''_{\mathbf{B}}$ and $\lambda = \max_i |q''_i|$,

$$\left| \frac{G(\mathbf{q}'_{\mathbf{B}} + \theta \mathbf{q}''_{\mathbf{B}}) - G(\mathbf{q}'_{\mathbf{B}})}{\theta} + \int_{-M}^{M} \frac{F(\mathbf{q}'_{\mathbf{B}} + \theta \mathbf{q}''_{\mathbf{B}} + t) - F(\mathbf{q}'_{\mathbf{B}} + t)}{\theta}\, dt \right|$$

$$\leqq 2\lambda \sum_{i=1}^{m} [1 - F^i(M - \lambda + q'_i) + F^i(-M + \lambda + q'_i)]. \tag{5.128}$$

The right-hand side of this inequality converges to zero as $M \to +\infty$, uniformly in θ. For each M the left-hand side converges to

$$\left| \lim_{\theta \to 0} \frac{G(\mathbf{q}'_{\mathbf{B}} + \theta \mathbf{q}''_{\mathbf{B}}) - G(\mathbf{q}'_{\mathbf{B}})}{\theta} + \int_{-M}^{M} \sum_{i=1}^{m} F_i(\mathbf{q}'_{\mathbf{B}} + t) q''_i\, dt \right|,$$

since F has a density and is therefore differentiable. This establishes that G is differentiable, with

$$\begin{aligned} G_i(\mathbf{q_B}) &= -\int_{-\infty}^{+\infty} F_i(\mathbf{q_B} + t)\, dt \\ &= -\int_{-\infty}^{+\infty} F_i(\mathbf{q_B} - q_i + t)\, dt \\ &= -\int_{\varepsilon_i = -\infty}^{+\infty} \int_{\varepsilon_2 = -\infty}^{\varepsilon_i - q_i + q_2} \cdots \int_{\varepsilon_m = -\infty}^{\varepsilon_i - q_i + q_m} f(\varepsilon_{\mathbf{B}})\, d\varepsilon_{\mathbf{B}} \\ &= -\operatorname{Prob}[\varepsilon_i - q_i \geqq \varepsilon_j - q_j \text{ for } j \in \mathbf{B}] \\ &= -P(i \mid \mathbf{B}, \mathbf{s}), \end{aligned} \tag{5.129}$$

with the second equality following by a change in the variable of integration from t to $t - q_i$ and the last inequality following from (5.2). Then (5.19) holds, and $\Sigma_{i=1}^{m} G_i(\mathbf{q_B}) = -1$. From (5.129) the mixed partial derivatives of G exist and are nonpositive and independent of order of differentiation. Thus we have established SS 5.1, SS 5.3, and SS 5.4.

The linear homogeneity of G in $(\mathbf{q_B}, \mathbf{r})$ follows from the linear homogeneity of the functions α_i in (5.12) which imply $F(\lambda\boldsymbol{\varepsilon}_{\mathbf{B}}, \lambda\mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s}) = F(\boldsymbol{\varepsilon}_{\mathbf{B}}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathbf{s})$ for $\lambda > 0$. The convexity of G follows by noting that $\Sigma_{i=1}^{M} G_i = -1$ and $G_{ij} \leqq 0$ for $i \neq j$ imply $G_{ii} = -\Sigma_{j \neq i} G_{ij} \geqq 0$. Hence the hessian of G has a weakly dominant positive diagonal. Then SS 5.2 holds.

To establish SS 5.5, note from (5.129) that

$$\lim_{q_i \to -\infty} G_i(\mathbf{q_B}) = -\lim_{q_i \to -\infty} \int_{\varepsilon_i = -\infty}^{+\infty} \int_{\varepsilon_2 = -\infty}^{\varepsilon_i - q_i + q_2} \cdots \int_{\varepsilon_m = -\infty}^{\varepsilon_i - q_i + q_m} f(\boldsymbol{\varepsilon}_{\mathbf{B}})\, d\boldsymbol{\varepsilon}_{\mathbf{B}}$$

$$= -\int_{\boldsymbol{\varepsilon}_{\mathbf{B}} = -\infty}^{+\infty} f(\boldsymbol{\varepsilon}_{\mathbf{B}})\, d\boldsymbol{\varepsilon}_{\mathbf{B}} = -1. \tag{5.130}$$

Finally SS 5.6 follows from RUM 5.1. Hence G is a social surplus function.

It is immediate from the definition (5.18) of $\bar{V}$, and the previously established condition (5.19), that $\bar{V}$ is a social indirect utility function, since $\bar{V}$ inverts to an expenditure function which is concave in prices.

Note finally that the conditions SS for a social surplus function imply immediately that the PCS system defined by (5.19) satisfies TPCS.

ii. Suppose G is a social surplus function satisfying SS. Then SS 5.4 and the condition $\Sigma_{i=1}^{m} G_i = -1$ implied by SS 5.3 establish that (5.19) defines a PCS system. It is immediate from SS that this PCS satisfies TPCS.

We next will establish the existence of an AIRUM form such that G satisfies (5.17). Define

$$F(\boldsymbol{\varepsilon}_{\mathbf{B}}) = \int_{-\infty}^{\varepsilon_1} \pi_1(0, \varepsilon_2 - t, \ldots, \varepsilon_m - t)\psi(t)\, dt, \tag{5.131}$$

where ψ is an arbitrary density. From TPCS 5.4 $\lim_{\varepsilon_i \to -\infty} F(\boldsymbol{\varepsilon}_{\mathbf{B}}) = 0$. Also, $\lim_{\boldsymbol{\varepsilon}_{\mathbf{B}} \to +\infty} \pi_1(0, \varepsilon_2, \ldots, \varepsilon_m) = 1$,

$$\lim_{\varepsilon_{\mathbf{B}}\to+\infty} F(\varepsilon_{\mathbf{B}}) = \int_{-\infty}^{+\infty} \pi_1(0,+\infty,\ldots,+\infty)\psi(t)\,dt = \int_{-\infty}^{+\infty} \psi(t)\,dt = 1.$$

From TPCS 5.5 and TPCS 5.6

$F_{1\ldots m}(\varepsilon_{\mathbf{B}}) = \pi_{1,2,\ldots,m}(0,\varepsilon_2-\varepsilon_1,\ldots,\varepsilon_m-\varepsilon_1)\psi(\varepsilon_1) \geqq 0.$

Hence (5.131) defines a cumulative distribution function characterizing an AIRUM form.

Consider the function

$$\tilde{G}(\mathbf{q}_{\mathbf{B}}) = \int_{-\infty}^{+\infty} [F(t+\mathbf{0}_{\mathbf{B}}) - F(t+\mathbf{q}_{\mathbf{B}})]\,dt$$

$$= -q_1 - \int_{-\infty}^{+\infty} [F(t+\mathbf{q}_{\mathbf{B}}-q_1) - F(t+\mathbf{0}_{\mathbf{B}})]\,dt, \qquad (5.132)$$

where the existence of $\tilde{G}$ and the second equality were established in the proof of part i. Then

$$\tilde{G}(\mathbf{q}_{\mathbf{B}}) = -q_1 - \int_{t=-\infty}^{+\infty}\int_{\tau=-\infty}^{t} [\pi_1(0,t+q_2-q_1-\tau,\ldots,t+q_m-q_1-\tau)$$

$$-\pi_1(0,t-\tau,\ldots,t-\tau)]\psi(\tau)\,d\tau\,dt$$

$$= -q_1 - \int_{\tau=-\infty}^{+\infty}\int_{t=\tau}^{+\infty} [\pi_1(0,q_2-q_1+t-\tau,\ldots,q_m-q_1+t-\tau)$$

$$-\pi_1(0,t-\tau,\ldots,t-\tau)]\psi(\tau)\,dt\,d\tau$$

$$= -q_1 - \int_{t=0}^{\infty} [\pi_1(0,q_2-q_1+t,\ldots,q_m-q_1+t)$$

$$-\pi_1(0,t,\ldots,t)]\,dt \int_{\tau=-\infty}^{+\infty} \psi(\tau)\,d\tau$$

$$= -q_1 - \int_{t=0}^{\infty} [\pi_1(-t, q_2 - q_1, \ldots, q_m - q_1)$$

$$- \pi_1(-t, 0, \ldots, 0)]\, dt. \tag{5.133}$$

Since $\pi_1 = -G_1$,

$$\begin{aligned}\tilde{G}(\mathbf{q_B}) &= -q_1 + [G(-t, 0, \ldots, 0) - G(-t, q_2 - q_1, \ldots, q_m - q_1)]_0^{+\infty} \\ &= -q_1 + G(0, q_2 - q_1, \ldots, q_m - q_1) = G(\mathbf{q_B}).\end{aligned} \tag{5.134}$$

Thus $\tilde{G}$ defined by (5.17) from this AIRUM form equals G.

iii. Suppose $\pi_i(\mathbf{q_B})$ is a PCS satisfying TPCS. Define $F(\boldsymbol{\varepsilon}_\mathbf{B})$ and $\tilde{G}(\mathbf{q_B})$ as in (5.131) and (5.132). Then F characterizes an AIRUM form, and

$$F(t + \mathbf{q_B} - q_1) = \int_{-\infty}^{t} \pi_1(0, t + q_2 - q_1 - \tau, \ldots, t + q_m - q_1 - \tau)\psi(\tau)\, d\tau$$

$$= \int_{-\infty}^{0} \pi_1(q_1, q_2 - \tau, \ldots, q_m - \tau)\psi(\tau + t)\, d\tau. \tag{5.135}$$

Hence

$$F_j(t + \mathbf{q_B} - q_1) = \int_{-\infty}^{0} \pi_{1,j}(q_1, q_2 - \tau, \ldots, q_m - \tau)\psi(\tau + t)\, d\tau, \tag{5.136}$$

implying from (5.132) that for $j > 1$

$$\tilde{G}_j(\mathbf{q_B}) = -\int_{t=-\infty}^{+\infty} \int_{\tau=-\infty}^{0} \pi_{1,j}(q_1, q_2 - \tau, \ldots, q_m - \tau)\psi(\tau + t)\, d\tau\, dt$$

$$= -\int_{\tau=-\infty}^{0} \pi_{1,j}(q_1 + \tau, q_2, \ldots, q_m) \int_{-\infty}^{+\infty} \psi(\tau + t)\, dt\, d\tau$$

$$= -\int_{\tau=-\infty}^{0} \pi_{j,1}(q_1 + \tau, q_2, \ldots, q_m)\, d\tau$$

$$= -\pi_j(\mathbf{q_B}) + \pi_j(-\infty, q_2, \ldots, q_m) = -\pi_j(\mathbf{q_B}) \qquad (5.137)$$

and

$$\tilde{G}_1(\mathbf{q_B}) = -1 - \sum_{j=2}^{m} \tilde{G}_j(\mathbf{q_B}) = -1 + \sum_{j=2}^{m} \pi_j(\mathbf{q_B}) = -\pi_1(\mathbf{q_B}). \qquad (5.138)$$

In (5.137) the third equality is a consequence of TPCS 5.6, and the last equality is a consequence of TPCS 5.4.

The foregoing argument establishes the existence of an AIRUM generating the PCS and of a function $\tilde{G}$, satisfying (5.17) and (5.18). The argument in part i then establishes that $\tilde{G}$ is a social surplus function satisfying SS and $\bar{V}$ defined by (5.18) is a social indirect utility function. This completes the proof of theorem 5.1.

To establish lemma 5.1, note that the existence of first moments of F implies

$$E \max_{i \in \mathbf{B}} (\varepsilon_i - q_i) = \int_{-\infty}^{+\infty} t \frac{d}{dt} F(\mathbf{q_B} + t)\, dt$$

exists, and hence

$$0 = \lim_{M \to \infty} \int_{-\infty}^{-M} t \frac{d}{dt} F(\mathbf{q_B} + t)\, dt \leqq \lim_{M \to \infty} [-MF(\mathbf{q_B} - M)] \leqq 0.$$

Similarly $\lim_{M \to \infty} M[1 - F(\mathbf{q_B} + M)] = 0$. Applying integration by parts to (5.17),

$$G(\mathbf{q_B}) = \lim_{M \to \infty} \int_{-M}^{M} [F(\mathbf{0_B} + t) - F(\mathbf{q_B} + t)]\, dt$$

$$= -\lim_{M \to \infty} t[F(\mathbf{q_B} + t) - F(\mathbf{0_B} + t)] \Big|_{-M}^{M}$$

$$+ \lim_{M \to \infty} \int_{-M}^{M} t \frac{d}{dt} [F(\mathbf{q_B} + t) - F(\mathbf{0_B} + t)]\, dt$$

$$= \int_{-\infty}^{+\infty} t \frac{d}{dt} F(\mathbf{q_B} + t)\, dt - \int_{-\infty}^{+\infty} t \frac{d}{dt} F(\mathbf{0_B} + t)\, dt$$

$$= E \max_{i \in \mathbf{B}} (\varepsilon_i - q_i) - E \max_{i \in \mathbf{B}} \varepsilon_i, \tag{5.139}$$

where the third equality follows from the existence of moments and the last equality from the definition of the expectation of the maximum component of a random vector. This proves lemma 5.1.

We next prove lemma 5.2. Suppose the first moments of F exist. Then, using the proof of lemma 5.1, and letting $G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) = E \max_{i \in \mathbf{B}} (\varepsilon_i - q_i)$, one has

$$G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) = G(\mathbf{0_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) + \int_{-\infty}^{+\infty} [F(t + \mathbf{0_B}) - F(t + \mathbf{q_B})]\, dt, \tag{5.140}$$

and the proof of theorem 5.1 implies that G satisfies SS.

Alternately suppose nonprice attributes are compensatable in the sense that, given $\mathbf{w_B}$ and $\delta > 0$, there exists $\theta > 0$, such that $F(\boldsymbol{\varepsilon}_\mathbf{B} + \theta, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) \geqq F(\boldsymbol{\varepsilon}_\mathbf{B}, \mathbf{r}, \mathbf{w'_B}, \mathbf{B}, \mathrm{s}) \geqq F(\boldsymbol{\varepsilon}_\mathbf{B} - \theta, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s})$ for all $\boldsymbol{\varepsilon}_\mathbf{B}$ and $|\mathbf{w'_B} - \mathbf{w_B}| < \delta$. Let $\bar{\mathbf{w}}_\mathbf{B}$ denote a fixed vector of nonprice attributes, and define

$$G(\mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s}) = \int_{-\infty}^{+\infty} [F(t + \mathbf{0_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}) - F(t + \mathbf{q_B}, \mathbf{r}, \mathbf{w_B}, \mathbf{B}, \mathrm{s})]\, dt. \tag{5.141}$$

Theorem 5.1 implies that $G(\mathbf{q_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s})$ exists and satisfies SS. Given $\mathbf{w'_B}$ with $|\mathbf{w'_B} - \mathbf{w_B}| < \delta$. one has

$$\begin{aligned} &F(t + \mathbf{0_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}) - F(t - \theta + \mathbf{q_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}) \\ &\quad \geqq F(t + \mathbf{0_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}) - F(t + \mathbf{q_B}, \mathbf{r}, \mathbf{w'_B}, \mathbf{B}, \mathrm{s}) \\ &\quad \geqq F(t + \mathbf{0_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}) - F(t + \theta + \mathbf{q_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathrm{s}). \end{aligned} \tag{5.142}$$

By dominated convergence $G(\mathbf{q_B}, \mathbf{r}, \mathbf{w'_B}, \mathbf{B}, \mathrm{s})$ exists, with

$$G(\mathbf{q_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathbf{s}) + \theta \geqq G(\mathbf{q_B}, \mathbf{r}, \mathbf{w}'_\mathbf{B}, \mathbf{B}, \mathbf{s}) \geqq G(\mathbf{q_B}, \mathbf{r}, \bar{\mathbf{w}}_\mathbf{B}, \mathbf{B}, \mathbf{s}) - \theta. \tag{5.143}$$

Then one can write

$$G(\mathbf{q_B}, \mathbf{r}, \mathbf{w}'_\mathbf{B}, \mathbf{B}, \mathbf{s}) = G(\mathbf{0_B}, \mathbf{r}, \mathbf{w}'_\mathbf{B}, \mathbf{B}, \mathbf{s}) + \int_{-\infty}^{+\infty} [F(t + \mathbf{0_B}, \mathbf{r}, \mathbf{w}'_\mathbf{B}, \mathbf{B}, \mathbf{s}) - F(t + \mathbf{q_B}, \mathbf{r}, \mathbf{w}'_\mathbf{B}, \mathbf{B}, \mathbf{s})]\, dt. \tag{5.144}$$

Since the integral in (5.144) has the form (5.17), the proof of theorem 5.1 establishes that G satisfies SS.

5.24 Appendix: The Elimination-by-Strategy Model

Tversky (1972b) has given examples showing that not all PCS satisfying RUM can be written as EBA models. However, one can establish an equivalence between models satisfying random preference maximization and a more general family of elimination models called elimination-by-strategy, EBS, models.

Let $\mathbf{H}$ be an abstract space of aspects, and $\mathscr{H}$ a σ-algebra of subsets of $\mathbf{H}$. Let $\mathbf{T}$ be a well-ordered set with an order relation $\prec$. Let $(\mathbf{K}, \mathscr{K})$ be a measurable space of functions from $\mathbf{T}$ into $\mathbf{H}$. Each $k \in \mathbf{K}$ is interpreted as a selection strategy.

Each alternative $i \in \mathbf{I}$ owns a set of aspects $\mathbf{D}_i \in \mathscr{H}$. Suppose a decision maker with strategy k and choice set $\mathbf{B}$ has remaining at $t \in \mathbf{T}$ a set $\mathbf{B}_t \subseteq \mathbf{B}$ of noneliminated alternatives. Then $i, j \in \mathbf{B}_t$ implies $k(t') \notin \mathbf{D}_i \Delta \mathbf{D}_j \equiv (\mathbf{D}_i - \mathbf{D}_j) \cup (\mathbf{D}_j - \mathbf{D}_i)$ for $t' \prec t$, and $i \in \mathbf{B}_t$, $j \in \mathbf{B} - \mathbf{B}_t$ implies $k(t') \in \mathbf{D}_i - \mathbf{D}_j$ for some $t' \prec t$ and $k(t'') \notin \mathbf{D}_i \Delta \mathbf{D}_j$ for $t'' \prec t'$. Since $\mathbf{T}$ is well ordered, $\mathbf{B}_t$ is well defined and monotonically nonincreasing to a nonempty limit set. A probability measure ν on $(\mathbf{K}, \mathscr{K})$ then determines choice probabilities, provided there is probability one of selecting a strategy for which the limit set of $\mathbf{B}_t$ is a singleton.

The EBS model is a random preference model: each $k \in \mathbf{K}$ corresponds to a lexicographic preference order on $\mathbf{H}$, and ν gives the distribution of preferences. To show that every random preference model can be rewritten as an EBS model, suppose $(\mathbf{J}, \mathscr{J}, \mu)$ is a probability space of preferences on

$\mathbf{I}$, and suppose that $\mathbf{T} = \mathbf{I}$, with $<$ ordering $\mathbf{I}$. Define $\mathbf{H} = \mathbf{J} \times \mathbf{I}$ and $\mathscr{H} = \mathscr{J} \otimes 2^{\mathbf{I}}$. Define $\mathbf{D}_i = \{(\succsim, l) \in \mathbf{J} \times \mathbf{I} \mid i \succsim l\}$ to be the set of aspects owned by i. Define $\mathbf{K}$ to be the set of functions $k : \mathbf{I} \to \mathbf{H}$ with $k(i) = (\succsim, i)$, where $\succsim \in \mathbf{J}$ is independent of i. Let $\succsim_k$ denote the preference relation in $\mathbf{J}$ determined by $k \in \mathbf{K}$, and define a measure v on $\mathbf{K}$ by $v(\mathbf{K}_1) = \mu(\{\succsim_k \in \mathbf{J} \mid k \in \mathbf{K}_1\})$ for each $\mathbf{K}_1 \in \mathscr{K}$. Given a choice set $\mathbf{B} \in \mathscr{B}$, suppose $k \in \mathbf{K}$ induces an elimination process that leads to choice of alternative $i \in \mathbf{B}$. Then for each $j \in \mathbf{B}, j \neq i$, there exists $l' \in \mathbf{I}$ such that $k(l') \in \mathbf{D}_i - \mathbf{D}_j$ and for $l'' \in \mathbf{I}$, $l'' < l'$, $k(l'') \notin \mathbf{D}_i \Delta \mathbf{D}_j$. This implies $i \succsim_k l'$ and not $j \succsim_k l'$, so that $i \succ_k j$. Let $\mathbf{K}_1 = \{k \in \mathbf{K} \mid i \succsim_k j$ for $j \in \mathbf{B}$, $j \neq i\}$. Then $v(\mathbf{K}_1) = \mu(\{\succsim \in \mathbf{J} \mid i \succsim j$ for $j \in \mathbf{B}, j \neq i\}) = P(i \mid \mathbf{B}, \mathbf{s})$, and the EBS model implies the same PCS as the random preference model.

Tversky's EBA model is a special case of the EBS model, corresponding to $\mathbf{K} = \mathbf{H}^{\infty}$ and v the product measure induced by independent sampling from a probability measure η on $\mathbf{H}$.

References

Amemiya, T. 1976. Specification and Estimation of a Multinomial Logit Model. Technical report 211. Institute of Mathematical Studies in the Social Sciences, Stanford University.

Amemiya, T. 1978. On a Two-Step Estimation of a Multivariate Logit Model. *Journal of Econometrics*. 8: 13–21.

Becker, G., and G. Stigler. 1977. De Gustubus Non Est Disputandum. *American Economic Review*. 67: 76–90.

Berndt, E., J. Hausman, B. Hall, and R. Hall. 1974. Estimation and Inference in Non-Linear Structural Models. *Annals of Economic and Social Measurement*. 3: 653–665.

Ben-Akiva, M., and Lerman, S. 1979. Disaggregate Travel and Mobility Choice Models and Measures of Accessibility. In *Behavior Travel Modelling*, ed. D. Hensher and P. Stopher. London: Coom Helm.

Block, H., and J. Marschak. 1960. Random Orderings and Stochastic Theories of Response. In *Contributions to Probability and Statistics*, ed. I. Olkin. Stanford, Calif.: Stanford University Press.

Bock, R., and L. Jones. 1968. *The Measurement and Prediction of Judgment and Choice*. San Francisco: Holden-Day.

Cardell, S. 1977. Multinomial Logit with Correlated Stochastic Terms. Working paper. Charles River Associates, Boston, Mass.

Chipman, J. 1974. Homothetic Preferences and Aggregation. *Journal of Economic Theory*. 8: 26–38.

Clark, C. 1961. The Greatest of a Finite Set of Random Variables. *Operations Research*. 9: 145–162.

Coombs, C. 1964. A Theory of Data. New York: Wiley.

Corbin, R., and A. Marley. 1974. Random Utility Models with Equality. *Journal of Mathematical Psychology*. 3: 174–193.

Cosslett, S. 1981. Maximum Likelihood Estimator for Choice-Based Samples. *Econometrica*. 9: 1289–1316.

Cox, D. 1970. *Analysis of Binary Data*. London: Methuen.

Daganzo, C., F. Bouthelier, and Y. Sheffi. 1977. Multinomial Probit and Qualitative Choice: A Computationally Efficient Algorithm. *Transportation Science*. 11: 338–358.

Daly, A., and S. Zachary. 1979. Improved Multiple Choice Models. In *Identifying and Measuring the Determinants of Mode Choice*. ed. D. Hensher and Q. Dalvi. London: Teakfield.

Debreu, G. 1960. Review of R. D. Luce *Individual Choice Behavior*. *American Economic Review*. 50: 186–188.

Debreu, G. 1962. New Concepts and Techniques for Equilibrium Analysis. *International Economic Review*. 3: 257–273.

Diewert, E. 1977. Generalized Slutsky Conditions for Aggregate Consumer Demand Functions. *Journal of Economic Theory*. 15: 353–362.

Edgell, S., W. Geisler, and J. Zinnes. 1973. A Note on a Paper by Rumelhart and Greeno. *Journal of Mathematical Psychology*. 10: 86–90.

Finney, D. 1964. *Statistical Method in Bioassay*. London: Griffin.

Fuss, M., D. McFadden, and Y. Mundlak. 1978. A Survey of Functional Forms in the Economic Analysis of Production. In *Production Economics; A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

Gaudry, M., and M. Dagenais. 1977. The DOGIT Model. Centre de Recherche sur les Transports, University of Montreal.

Gorman, W. 1953. Community Preference Fields. *Econometrica*. 21: 63–80.

Hausman, J., and D. A. Wise. 1978. A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*. 46: 403–426.

Hausman, J. 1978. Alternative Specifications of Probit Models of Qualitative Choice. Mimeographed Massachusetts Institute of Technology, Cambridge.

Heckman, J. 1978. Simple Statistical Models for Discrete Panel Data. *Annales de L'Insee*. 30–31: 227–270.

Horowitz, J. 1979. A Note on the Accuracy of the Clark Approximation for the Multinomial Probit Model. Mimeographed. Massachusetts Institute of Technology, Cambridge.

Krantz, D. 1974. Measurement Theory and Qualitative Laws in Psychophysics. In *Measurement, Psychophysics, and Neural Information Processing*, vol. 2, ed. D. Krantz et al. San Francisco: Freeman.

Krantz, D., D. Luce, P. Suppes, and A. Tversky. 1971. *Foundations of Measurement*, vol. 1: *Additive and Polynomial Representations*. New York: Academic Press.

Lau, L. 1977. Existence Conditions for Aggregate Demand Functions. Technical report 248. Institute for Mathematical Studies in the Social Sciences, Stanford University.

Luce, R. D. 1959. *Individual Choice Behavior; A Theoretical Analysis*. New York: Wiley.

Luce, R. D. 1977. The Choice Axiom after Twenty Years, *Journal of Mathematical Psychology*, 15: 215–233.

Luce, R. D., and P. Suppes. 1965. Preference, Utility, and Subjective Probability. In *Handbook of Mathematical Psychology*, vol. 3, ed. R. D. Luce, R. Bush, and E. Galanter. New York: Wiley.

Marschak, J. 1960. Binary-Choice Constraints and Random Utility Indicators. In *Mathematical Methods in the Social Sciences*. ed. K. Arrow, S. Karlin, and P. Suppes. Stanford: Stanford University Press.

McFadden, D. 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1974. The Measurement of Urban Travel Demand. *Journal of Public Economics*. 3: 303–328.

McFadden, D. 1975. On Independence, Structure, and Simultaneity in Transportation Demand Analysis. Working paper 7511. Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley.

McFadden, D. 1976. Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement*. 5: 363–390.

McFadden, D. 1978a. Modelling the Choice of Residential Location. In *Spatial Interaction Theory and Residential Location*, ed. A Karlgvist et al. North Holland: Ansterdam, pp. 75–96.

McFadden, D. 1978b. Cost, Revenue, and Profit Functions. In *Production Economics; A Dual Approach to Theory and Applications*, vol. 1, ed. M. Fuss and D. McFadden. Amsterdam: North-Holland.

McFadden, D., and M. Richter. 1970. Revealed Stochastic Preference. Mimeographed. Department of Economics, Massachusetts Institute of Technology.

McFadden, D., A. Talvitie, and Associates. 1977. *Demand Model Estimation and Validation*. Urban travel demand forecasting project, final report, vol. 5. Institute of Transportation Studies, University of California, Berkeley.

McFadden, D., W. Tye, and K. Train. 1977. An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model. *Transportation Research Record*. 637: 39–45.

McKenzie, L. 1957. Demand Theory without a Utility Index. *Review of Economic Studies*. 24: 185–189.

McLynn, J. 1973. A Technical Note on a Class of Fully Competitive Modal Choice Models. DIM Corporation, Bethesda, Md.

Muellbauer, J. 1976. Community Preferences and the Representative Consumer. *Econometrica*. 44: 979–999.

Prais, S., and H. Houthakker. 1971. *The Analysis of Family Budgets*. Cambridge, England: Cambridge University Press.

Quandt, R. 1968. Estimation of Modal Splits. *Transportation Research*. 2: 41–50.

Rosen, H., and K. Small. 1979. Applied Welfare Economics with Discrete Choice Models. Working paper 319. National Burean of Economic Research, Cambridge, Mass.

Rumelhart, D., and J. Greeno. 1971. Similarity between Stimuli: An Experimental Test of the Luce and Restle Choice Models. *Journal of Mathematical Psychology*. 8: 370–381.

Sattath, S., and A. Tversky. 1977. Additive Similarity Trees. *Psychometrika*. 42: 319–345.

Shapiro, P. 1975. Aggregation, Social Choice, and the Existence of a Social Utility Function. Working paper 55. U. S. Bureau of Labor Statistics, Washington, D.C.

Simon, H. 1977. On How to Decide What to Do. C.I.P. working paper. Carnegie-Mellon University, Pittsburgh.

Sjöberg, L. 1977. Choice Frequency and Similarity. *Scandinavian Journal of Psychology*. 18: 103–115.

Thurstone, L. 1927. A Law of Comparative Judgment. *Psychological Review*. 34: 273–286.

Tobin, J. 1958. Estimation of Relationships for Limited Dependent Variables. *Econometrica*. 26.

Train, K. 1978. A Validation Test of a Disaggregate Mode Choice Model. *Transportation Research*. 12: 167–174.

Train, K., and D. McFadden 1978. The Goods-Leisure Tradeoff and Disaggregate Work Trip Mode Choice Models. *Transportation Research;* 12: 349–353.

Tversky, A. 1972a. Elimination-by-Aspects: A Theory of Choice. *Psychological Review*. 79: 281–299.

Tversky, A. 1972b. Choice-by-Elimination. *Journal of Mathematical Psychology*. 9: 341–367.

Tversky, A., and S. Sattath. 1979. Preference Trees. *Psychological Review*. 86: 542–573.

Warner, S. 1962. *Stochastic Choice of Mode in Urban Travel; A Study in Binary Choice*. Evanston: Northwestern University Press.

Williams, H. 1977. On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit. *Environment Planning*. A.9: 285–344.

Yellot, J. 1977. The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. *Journal of Mathematical Psychology*. 15: 109–144.